Probability Distributions

- General concepts
  1. Random variable
  2. Parameters
  3. Mean, variance of a random variable
- Bernoulli/Binomial
  1. Gives probability of k 1's in n (independent) trials for a binary (0/1) random variable
- Poisson
  1. Gives probability of counts per unit time/area/vol.
- Normal
  1. Continuous RV's
  2. Standard normal
  3. Probability, Quantile calculations
  4. Can approximate Binomial/Poisson

Probability Distribution

Definition: A **random variable** is a characteristic whose values arise as a result of chance factors.

Definition: A **probability distribution** gives the probability of obtaining all possible (sets of) values of a random variable. It gives the probability of the outcomes of an experiment. Note that a probability distribution is an example of the "classical" definition of probability. Think of it as a histogram of the entire population.

Population	↔	Sample
Random variable	↔	Measurement
Probability dist.	↔	Frequency dist.
Parameters	↔	Statistics (Estimates)

Probability Distributions

Used to provide a mathematical description of the relative likelihoods of various possible outcomes of an experiment. It is a **model** of reality, not reality!

A. Discrete variables
   1. Binomial - sums of 0/1 outcomes
      - underlies many epidemiologic applications
      - basic model for logistic regression
   2. Poisson - counts of events, rates
      - basic model for log-linear analysis

B. Continuous variables
   1. Normal - bell-shaped curve; many measurements are approximately normally distributed.
   2. t- distribution
   3. Chi-square distribution (χ^2)

Probability Distributions - Discrete

A discrete random variable, X, has a countable (e.g. 1, 2, 3, …) number of values. The probability distribution of X lists the possible values of X and their probabilities:

Value of X:	x_1	x_2	x_3	x_4	…
Probability:	p_1	p_2	p_3	p_4	…

The probabilities must satisfy the following:

1) Each p_i is between 0 and 1.

2) The sum of the p_i's is equal to 1.

Probability Distributions - Discrete

E.g. Household size

Value of X = household size	1	2	3	4	5	6	7
P(X)	.20	.32	.18	.15	.07	.03	

What is the missing value of P(X) for X = 7 ?

Mean and Variance of a Random Variable

Suppose we interview 10 households and find the number of people living in each. Here are the data:

2 1 4 3 4 1 2 2 7 1

How do we compute the mean number in this sample?

$$\begin{aligned}\bar{X} &= \frac{2+1+4+3+4+1+2+2+7+1}{10} \\ &= \tfrac{3}{10}*1+\tfrac{3}{10}*2+\tfrac{1}{10}*3+\tfrac{2}{10}*4+\tfrac{1}{10}*7 \\ &= 2.7\end{aligned}$$

- Notice that the sample mean has the form $\bar{X} = \sum_x f_x x$ where f_x is the relative frequency of the value x (i.e. weighted mean).

Mean and Variance of a Random Variable

We can extend the concept of a weighted mean to define the mean of a random variable that follows a known probability distribution. The mean of a random variable X is denoted by E(X) (often, the symbol μ is used). E(X) should be thought of as the mean of the population.

Suppose the probability distribution of a discrete random variable X is given by P(x) (= p_x). Then

$$\mu = \mathrm{E(X)} = \sum_x p_x x$$

Also, we will use V(X) (or the symbol σ^2) to denote the population variance

$$\sigma^2 = \mathrm{V(X)} = \mathrm{E}\left[(\mathrm{X}\text{-}\mu)^2 \right]$$

Means and Variance of the Sum of independent RV's

If X_1, X_2, …,X_n are (any) **independent** random variables and if we define Y= X_1+ X_2+ …+X_n

1. Means add:

$$E[Y]= E[X_1]+E[X_2]+ \ldots +E[X_n]$$

2. Variances add:

$$V[Y]= V[X_1]+V[X_2]+ \ldots +V[X_n]$$

Bernoulli Distribution

A Bernoulli RV is a discrete RV with only 2 possible outcomes, which we denote by 0 or 1 (e.g. coin toss)

Assumptions:

1. Two possible outcomes – "success" (1) or failure (0).
2. The probability of success = p; same for each trial.
3. The outcome of one trial has no influence on later outcomes (independent trials).

E.g.

Coin toss p = 0.5

Roll a 1 on a die p = 1/6

Survive 5 years following breast cancer diagnosis p = .85

Bernoulli RV - Mean and Variance

Consider a Bernoulli random variable with success probability **p**.

P[X=1] = p

P[X=0]=1-p

Mean: $\mu = E[X] = p$

Variance: $\sigma^2 = V[X] = p(1-p)$

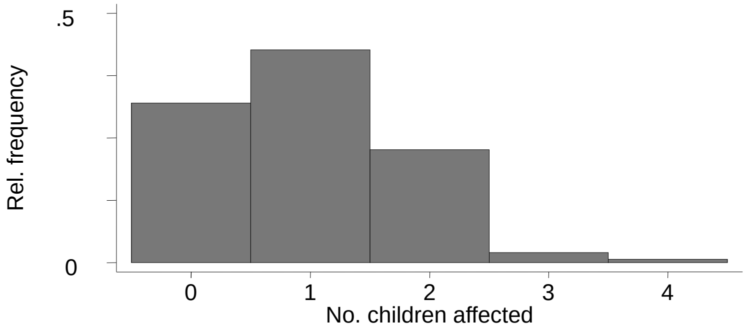
Binomial Distribution - Motivation

Question: In a family where both parents are carriers for a recessive trait, what is the probability that in a family of 4 children exactly 1 child would be affected?

- ➢ There are three key bits of information in this question:
  - probability given child is affected (recessive ⇒ p = .25)
  - number of children (n = 4)
  - number affected (k = 1)
- ➢ p, n, k can change –we need a general approach to address …
  - What is the probability that 2 would be affected?
  - In a family of 6 children, what is the probability that 1 child is affected?
  - What if the trait is dominant?

Binomial Distribution - Motivation

We might think of taking a large sample of families with 4 children and both parents carriers of the recessive trait. Then plot a barchart of the relative frequencies of the number of children affected. But that is hard to do … and still just a sample.



Relative frequency = Probability

Binomial Random Variable

A binomial random variable is simply the total number of successes in **n** Bernoulli trials. That is, a binomial random variable is the sum of **n** Bernoulli trials.

Example: number of affected children in a family of 4.

What we need to know is:

1. How many ways are there to get k "successes" (k=0,…4) in n trials?
2. What's the probability of any given outcome with exactly k successes (does order matter)?

How many ways are there of getting k "successes" in n trials (if order doesn't matter)? The formula is

$$C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

E.g. How many ways of getting 3 affected's (affected = 1) in 4 children?

Child number			
1	2	3	4
1	1	1	1
1	1	1	0
1	1	0	1
0	1	1	1
1	0	1	1
1	1	0	0
1	0	1	0
1	0	0	1
0	1	0	1
0	1	1	0
0	0	1	1
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	0	0	0

Q: What is the probability of any one of these outcomes? say

1	1	1	0
p	p	p	(1-p)

A: Independent, so multiply ….

$P(1110) = p^3(1-p)$

- In general, any <u>one</u> sequence of k pluses and (n-k) minuses will have probability

  $p^k(1-p)^{n-k}$

- But there are $\binom{n}{k}$ such sequences, so ...

<u>Binomial Probabilities</u>

What is the probability that a binomial random variable with **n** trials and success probability **p** will yield exactly **k** successes?

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

This is the **probability distribution function (or probability mass function)** for the binomial distribution.

Assumptions:

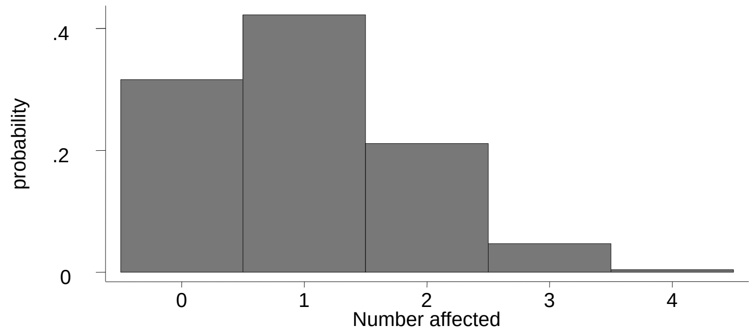
1) Two possible outcomes - success (1) or failure (0) - for each of n trials.

2) The probability of success, p, is the same for each trial.

3) The outcome of one trial has no influence on later outcomes (independent trials).

4) The random variable of interest is the total number of successes.

n=4, p = .25



This looks like the barchart I showed before and the interpretation is similar, but this is not a relative frequency plot for a sample of data. Rather it gives the probability distribution for the entire population. **It is derived theoretically, not from data.**

Binomial Probabilities - Example

Returning to the original question: What is the probability of exactly 1 affected child in a family of 4? (recessive trait, carrier parents)

Mean and Variance - Binomial

Recall that a binomial RV is just the **sum** of **n** independent Bernoulli random variables.

$Y \sim \text{Binomial}(n,p)$

$X_1, \ldots, X_n \sim \text{Bernoulli}(p)$

$Y = X_1 + \ldots + X_n$

So

$$\begin{aligned} E(Y) &= E(X_1) + \ldots + E(X_n) \\ &= p + \ldots + p \\ &= np \end{aligned}$$

$$\begin{aligned} V(Y) &= V(X_1) + \ldots + V(X_n) \\ &= p(1-p) + \ldots + p(1-p) \\ &= np(1-p) \end{aligned}$$

Poisson Distribution

Often data are collected on the number of occurrences of a particular event.

- Emergency room admissions during the week January 1 - 7, 1996.
- Number of bacterial colonies on an agar plate.
- Number of incident cases of colon cancer among males 30-35 during 1996.
- Number of cycling related head injuries in Seattle during January.

What are some common features of these examples?

Poisson Distribution - Assumptions

The **Poisson** distribution is a common model for counts, particularly of "rare" events. Such data take only integer values and the probability of "large" counts declines rapidly.

If the following assumptions are met, then the Poisson would be an appropriate model:

1. For a small time interval (or area), Δt, the following properties hold:
   (a) $P[X=1] \approx \lambda \, \Delta t$
   (b) $P[X>1] \approx 0$
2. **Stationarity**: The rate of event occurrence is constant throughout the time interval.
3. **Independence**: Events occurring in one time interval do not change the probability for events in another time interval.

Poisson Distribution - Properties

The probability of k events occurring in a time period t for a Poisson random variable with rate parameter λ is:

$$P[X = k] = \frac{e^{-\mu} \mu^k}{k!}$$

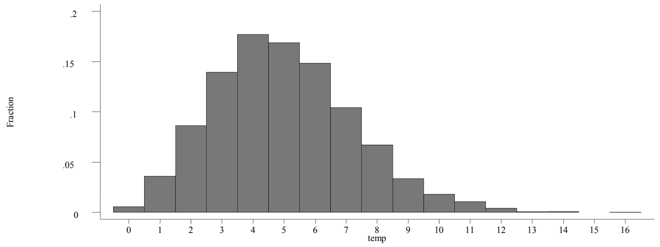
where $\mu = \lambda \times t$.

MEAN: $E[X] = \mu$.

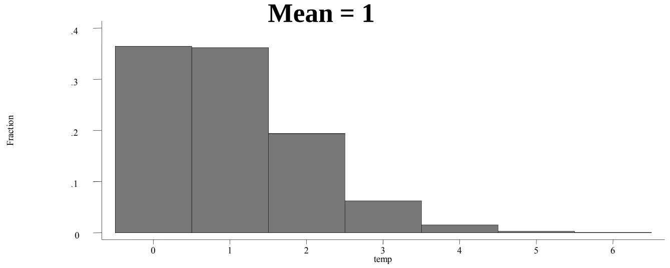
VARIANCE: $V[X] = \mu$.

If $X_1, X_2, \ldots, X_n$ are **independent** Poisson random variables with parameters μ_1, μ_2, etc., then $Y = X_1 + X_2 + \ldots + X_n$ is also Poisson with parameter $\mu_1 + \mu_2 + \ldots + \mu_n$.

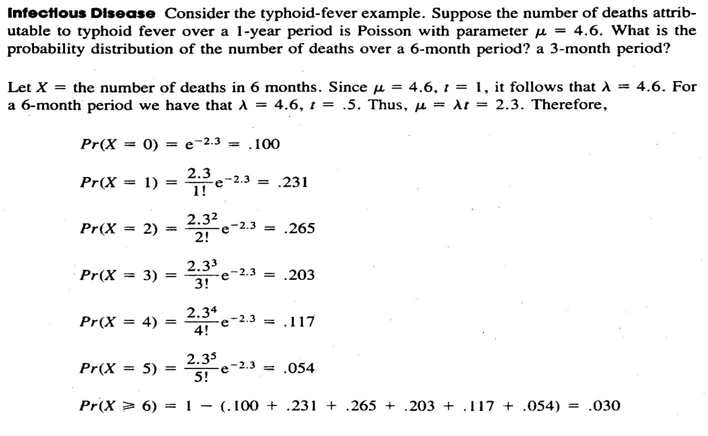
Mean = 5



Mean = 1



Poisson Distribution - Example



Infectious Disease Consider the typhoid-fever example. Suppose the number of deaths attributable to typhoid fever over a 1-year period is Poisson with parameter $\mu = 4.6$. What is the probability distribution of the number of deaths over a 6-month period? a 3-month period?

Let X = the number of deaths in 6 months. Since $\mu = 4.6$, $t = 1$, it follows that $\lambda = 4.6$. For a 6-month period we have that $\lambda = 4.6$, $t = .5$. Thus, $\mu = \lambda t = 2.3$. Therefore,

$$Pr(X = 0) = e^{-2.3} = .100$$

$$Pr(X = 1) = \frac{2.3}{1!}e^{-2.3} = .231$$

$$Pr(X = 2) = \frac{2.3^2}{2!}e^{-2.3} = .265$$

$$Pr(X = 3) = \frac{2.3^3}{3!}e^{-2.3} = .203$$

$$Pr(X = 4) = \frac{2.3^4}{4!}e^{-2.3} = .117$$

$$Pr(X = 5) = \frac{2.3^5}{5!}e^{-2.3} = .054$$

$$Pr(X \geqslant 6) = 1 - (.100 + .231 + .265 + .203 + .117 + .054) = .030$$

- ➢What is the probability of exactly 1 death due to typhoid in a 6 mo. period?
- ➢What is the probability of one or more deaths due to typhoid in a 6 mo. period?

Poisson Distribution - Warning

- My experience has been that the Poisson distribution is seldom adequate as a model for counts because the variance of the counts is often much larger than the mean
- When this happens inferential procedures (i.e. hypothesis tests and confidence intervals) based on a Poisson model tend to go disastrously wrong

Examples:

Number of episodes of GUD over the course of 18 mo. followup

Number of sex acts in a 3 mo. period in heterosexual African women

Number of sex partners in a 3 mo. period in MSM in Peru

Binomial versus Poisson

Binomial

1. Discrete, bounded
2. Parameters - **n,p**
3. Sum of n independent 0/1 outcomes
4. Sample proportions, logistic regression

Poisson

1. Discrete, unbounded
2. Parameters - $\mathbf{t,\lambda}$ or just $\boldsymbol{\mu}=t \times \lambda$
3. Independent time (area) intervals, constant rate
4. Sample counts, log-linear regression

Probability Distributions

- **General concepts**
  1. Random variable
  2. Parameters
  3. Mean, variance of a random variable
- **Bernoulli/Binomial**
  1. Gives probability of k 1's in n (independent) trials for a binary (0/1) random variable
- **Poisson**
  1. Gives probability of counts per unit time/area/vol.
- **Normal**
  1. Continuous RV's
  2. Standard normal
  3. Probability, Quantile calculations
  4. Can approximate Binomial/Poisson

Continuous Distributions

A **continuous random variable**, X, takes on all possible values in an interval. The probability distribution of X is described by a **probability density curve** (remember?).

For continuous variables we talk about the probability that the random variable falls within an interval:

$$P[\text{weight} = 70.000\text{kg}] \approx 0$$

$$P[69.0\text{kg} \leq \text{weight} \leq 71.0\text{kg}] = 0.08$$

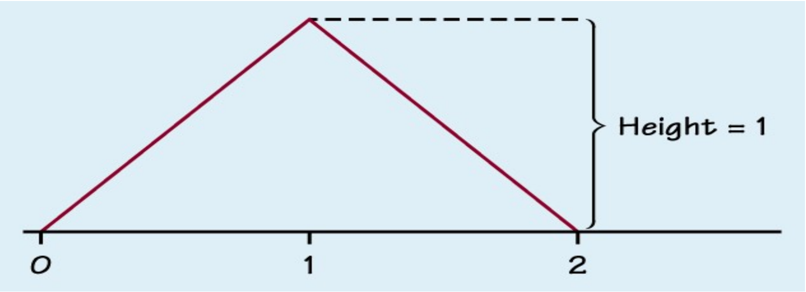
The probability that X falls in an interval is the <u>area</u> under the density curve within that interval.

Recall that the key properties of a density curve, f(x), are

1) $f(x) \geq 0$

2) Total area under the function f(x) is 1.0.

E.g. Suppose that X can vary from 0 to 2 and has a triangular distribution:



- ➢Is the area under the curve equal to 1?
- ➢What is the probability X is less than 1.0?
- ➢What is the probability X is less than 0.5?

Normal Distribution

Recall that one example of a continuous distribution is the **normal distribution.** Its probability density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} \right)$$

where

$\pi \approx 3.14$ (a constant)

The normal distribution has two parameters:

μ = the mean of X
σ = the standard deviation of X

- ➢ We can interpret μ and σ as the mean and standard deviation of the population (whereas $\bar{x}$ and s are the mean and standard deviation of the sample).
- ➢ The normal distribution is a very common **model** for data.

Normal Distribution - Calculating Probabilities

Example: Serum cholesterol is approximately normally distributed with mean 219 mg/mL and standard deviation 50 mg/mL. If the clinically desirable range is < 200 mg/mL, then what proportion of the population falls in this range?

X = serum cholesterol in an individual.

μ =

σ =

$$P[x < 200] = \int_{-\infty}^{200} \frac{1}{50\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x-219)^2}{50^2} \right) dx$$

negative values for cholesterol - huh?

Standard Normal Distribution - Calculating Probabilities

First, let's consider the **standard normal** - N(0,1). We will usually use Z to denote a random variable with a standard normal distribution. The density of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2 \right)$$

and the **cumulative distribution** of Z (i.e. P(Z < z)) is:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x^2 \right) dx$$

Tables (BM Table B) and computer routines are available for calculating these probabilities.

Table entry for z is the area under the standard Normal curve to the left of z.

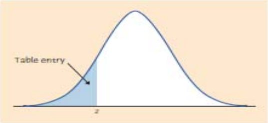


TABLE A Standard Normal cumulative proportions

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Table entry for z is the area under the standard Normal curve to the left of z.

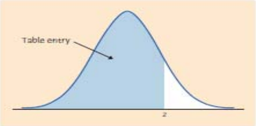


TABLE A Standard Normal cumulative proportions *(continued)*

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Standard Normal Probabilities

Using BM table B, find

$P[Z \le 1.65] =$

$P[Z \ge 0.5] =$

$P[-1.96 \le Z \le 1.96] =$

$P[-0.5 \le Z \le 2.0] =$

Converting to Standard Normal

Q: This solves the problem for the N(0,1) case. Do we need a special table for every (μ,σ)?

A: No! Any normal random variable can be **transformed** to N(0,1)

Standardize:

$$Z = \frac{X - \mu}{\sigma}$$

$$Z \sim N(0,1)$$

Linear transformations of Normal RV

- If $X \sim N(\mu, \sigma^2)$ and $Z = aX + b$ then

  $E(Z) = a\mu + b$
  $V(Z) = a^2\sigma^2$

  $Z \sim N(a\mu + b, a^2\sigma^2)$

- If we let $a = 1/\sigma$ and $b = \mu/\sigma$ we get the result on the previous page

Linear functions of normal RV's are also normal.

Normal Distribution - Calculating Probabilities

Return to cholesterol example...

Serum cholesterol is approximately normally distributed with mean 219 mg/mL and standard deviation 50 mg/mL. If the clinically desirable range is < 200 mg/mL, then what proportion of the population falls in this range?

$$
\begin{aligned}
P(X < 200) &= P\left(\frac{X-\mu}{\sigma} < \frac{200-219}{50}\right) \\
&= P\left(Z < \frac{200-219}{50}\right) \\
&= P(Z < -0.38) \\
&=
\end{aligned}
$$

Suppose X~ N(0,1)

$P(X < 1.96) =$

$P(-1.65 < X \le 1.65) =$

$P(-1.0 < X < .5) =$

$P(1.0 > X > 0) =$

$P(|X| > 2.0) =$

$P(|X| \le 1.0) =$

Suppose X ~ $N(2,2^2)$

$P(X < 0) =$

$P(X > 4) =$

$P(0 < X < 4) =$

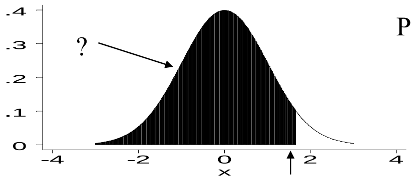
$P(|X| > 2) =$

Normal Quantiles

Go back to BM, table B ….

Notice that we can use the table two ways:

1) Given a particular x value (the quantile) we can look up the probability:

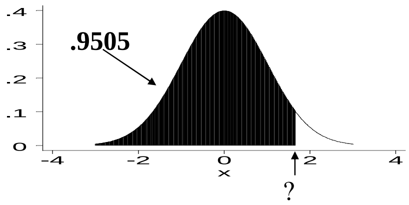


$P(Z \le \mathbf{1.65}) = ?$

1.65

z	.00	.01	.02	.03	.04	.05 ...
:	:	:	:	:		
1.6	.9452	.9463	9574	.9484	.9495	.9505 ...
:	:	:	:	:	:	:

2) Given a particular probability, we can look up the quantile:



$P(Z \leq ?) =$ **.9505**

z	.00	.01	.02	.03	.04	.05 ...
:	:	:	:	:		
1.6	.9452	.9463	9574	.9484	.9495	**.9505** ...
:	:	:	:	:	:	:

➢ z_α is the value such that $P(Z \leq z_\alpha) = \alpha$

Eg: $z_{.05} = -1.65$
$z_{.95} = 1.65$
$z_{.975} = 1.96$

Notice that $z_\alpha = -z_{1-\alpha}$

Supplementary Problems

Suppose $X \sim N(0,1)$

$P(X < ?) = .95$

$P(X < ?) = .167$

$P(X > ?) = .5$

$P(-1.96 < X < ?) = .95$

Challenge: Suppose $X \sim N(2,2^2)$ … find $P(X < ?) = .167$

Normal Approximation to Binomial and Poisson

An interesting property of the normal distribution is that it can be used to approximate other distributions under certain circumstances.

Example 1

Suppose the number of bicycling accidents in Seattle is Poisson with 84.5 accidents/year. What is the probability of observing more than 100 bicycle accidents in 1996?

Random variable?

Distribution?

Parameter(s)?

Question?

Normal Approximation to Binomial and Poisson

Example 2

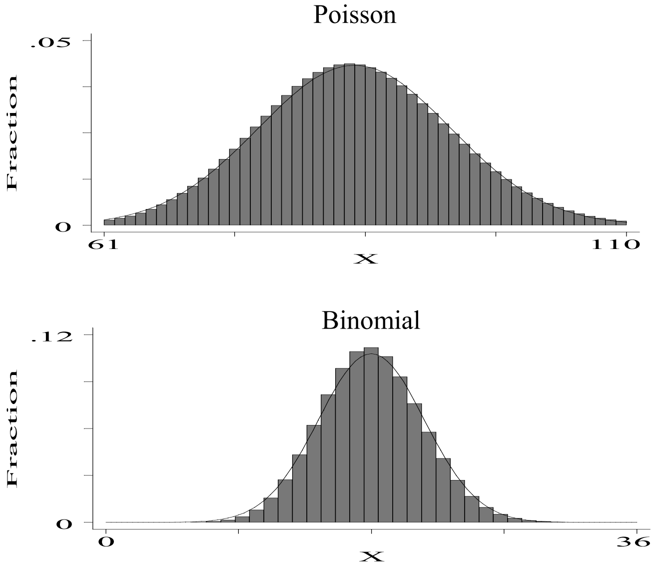
Suppose the prevalence of HPV in women 18 -22 years old is 0.30. What is the probability that in a sample of 60 women from this population less than 10 would be infected?

Random variable?

Distribution?

Parameter(s)?

Question?



Normal Approximation to Binomial and Poisson

1. **Binomial**
   - When both **np** and **n(1-p)** are "large" (both ≥ 10) the normal may be used to approximate the binomial.
   - X ~ bin(n,p)

     E(X) = np
     V(X) = np(1-p)
   - X is approximately N(np,np(1-p))
2. **Poisson**
   - When μ is "large" ($\mu \geq 10$) the normal may be used to approximate the Poisson.
   - X ~ Po(μ)

     E(X) = μ
     V(X) = μ
   - X is approximately N(μ, μ)

Normal Approximation to Binomial and Poisson

Example 1

Suppose the number of bicycling accidents in Seattle is Poisson with 84.5 accidents/year. What is the probability of observing more than 100 bicycle accidents in 1996?

Random variable?

⇒ X = number of bicycle accidents

Distribution?

⇒ Poisson

Parameter(s)?

⇒ $\mu = 84.5$

Question?

⇒ $P(X > 100) =$

⇒ normal approx. =

Normal Approximation to Binomial and Poisson

Example 2

Suppose the prevalence of HPV in women 18 -22 years old is 0.30. What is the probability that in a sample of 60 women from this population less than 10 would be infected?

Random variable?

⇒ X = number infected out of 60

Distribution?

⇒ Binomial

Parameter(s)?

⇒ $n = 60, \ p = .30$

Question?

⇒ $P(X \leq 9) =$

⇒ normal approx. =

Summary

- General concepts
  1. Random variable
  2. Parameters
  3. Mean, variance of a random variable
- Bernoulli/Binomial
  1. Gives probability of k 1's in n (independent) trials for a binary (0/1) random variable
- Poisson
  1. Gives probability of counts per unit time/area/vol.
- Normal
  1. Continuous RV's
  2. Standard normal
  3. Probability, Quantile calculations
  4. Can approximate Binomial/Poisson

Supplementary Problems

Suppose $X \sim N(0,1)$

$P(X < 1.96) = .975$

$P(-1.65 < X \leq 1.65) = .901$

$P(-1.0 < X < .5) = .5328$

$P(1.0 > X > 0) = .3413$

$P(|X| > 2.0) = P(X < -2) + P(X > 2) = .0456$

$P(|X| \leq 1.0) = P(-1 < X < 1) = .6826$

Suppose $X \sim N(2,2^2)$

$P(X < 0) = P(Z < (0-2)/2) = .1587$

$P(X > 4) = P(Z > (4-2)/2) = .1587$

$P(0 < X < 4) = P(-1 < Z < 1) = .6826$

$P(|X| > 2) = P(X < -2) + P(X > 2) = P(Z < -2) + P(Z > 0) = .5228$

Supplementary Problems

Suppose $X \sim N(0,1)$

$P(X < 1.645) = .95$

$P(X < -.97) = .166$

$P(X > 0) = .5$

$P(-1.96 < X < 1.96) =.95$

Suppose $X \sim N(2,2^2)$.

Note that $Z = (X-2)/2 \sim N(0,1)$.

$.167 = P(Z < -.97) = P((X-2)/2 < -.97) = P(X < .06)$

Sampling Distributions

- Summary statistics are random variables
- Linear combinations of normals → normal
- Central Limit Theorem: mean (sums) of non-normal data → normal
- $\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Variability of Summary Statistics

- ➢Individual observations are variable ….
- ➢The sample mean is the sum of individual observations …
- ➢Doesn't that mean that the sample mean is variable as well??

YES!

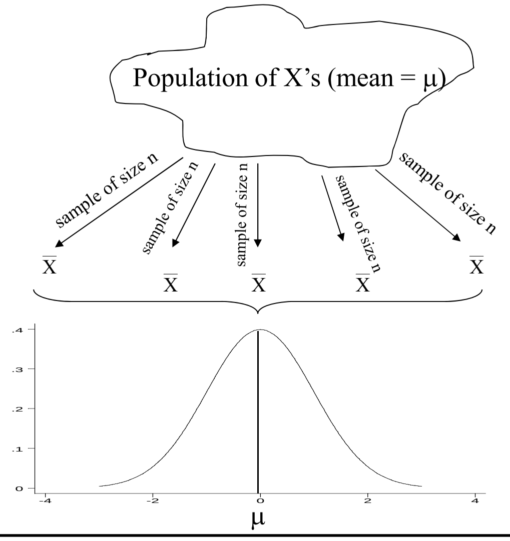
The sample mean is variable in the following sense: if I were to repeat the experiment (take a new sample) I would get a different mean.

This is a critically important concept!!!

And it applies to all summary statistics (sample variance, sample correlation, …)

The sampling distribution of a statistic is the distribution of the values of that statistic taken over all possible samples (of size n) from the population.

Distribution of the Sample Mean



Sums of Normal Random Variables

We already know that linear functions of a normal rv are normal. What about combinations (eg. sums) of normals?

If $X_j \sim N(\mu, \sigma^2)$ (indep) and $Y=\sum_{j=1}^{n} X_j$ then

$$Y \sim N\left(\sum_{j=1}^{n}\mu, \sum_{j=1}^{n}\sigma^2\right)$$

Combine this with what we have learned about linear functions of means and variances to get ...

Distribution of the Sample Mean

- When sampling from a **normally** distributed population (i.e. $X \sim N(\mu,\sigma^2)$)

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Sample mean is less variable than the individual observations

- When the population is **non-normal** but the sample size is large, the **Central Limit Theorem** applies.

Central Limit Theorem

Given a population with any non-normally distributed variables with a mean μ and a variance σ^2, then for large enough sample sizes, the distribution of the sample mean, $\overline{X}$ will be **approximately normal** with mean μ and variance σ^2/n.

$$\text{n large} \rightarrow \overline{\text{X}} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- In general, this applies for $n \geq 30$.
- As n increases, the normal approximation improves.

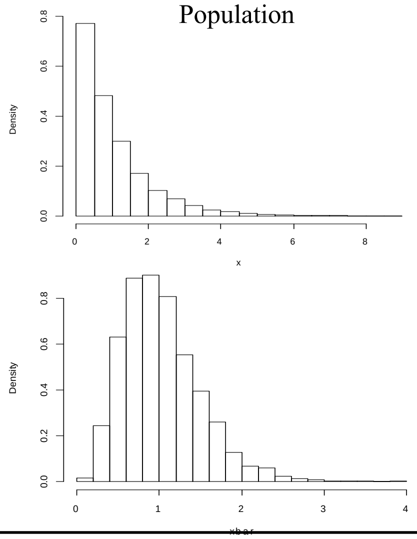
Eg. Sampling Distribution of the sample proportion

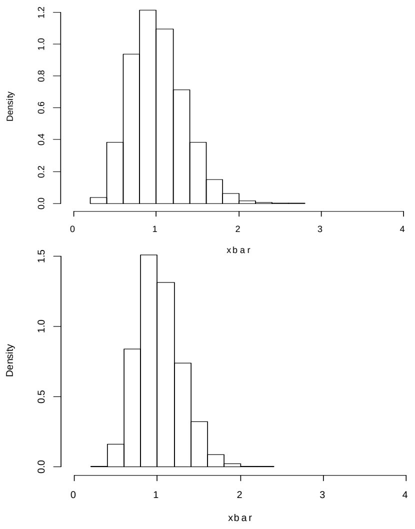
Assume $X_i \sim$ Bernoulli(p)

$$\hat{p} = \sum X_i / n$$

CLT ⇒ $\hat{p} \sim N(p, p(1-p)/n)$ for n large

Central Limit Theorem - Illustration





Distribution of Sample Mean

Example:

Suppose IQ scores are normally distributed with mean 100 and standard deviation 15.

a) What is the probability an individual has an IQ greater than 110?

$$P(X > 110) = P(Z > \frac{110-100}{15}) = P(Z > 0.67) = .2514$$

b) What is the probability that 10 randomly selected individuals have a mean IQ greater than 110?

$$P(\overline{X} > 110) = P(Z > \frac{110-100}{15/\sqrt{10}}) = P(Z > 2.1) = .0179$$

Summary

- Summary statistics are random variables
- The distribution of a summary statistic is called the sampling distribution of the statistic
- Linear combinations of normals → normal
- Central Limit Theorem: mean (sums) of non-normal data → normal
- σ = std. dev. of X; $\sigma / \sqrt{n}$ = std. dev. of $\overline{X}$
- $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Statistical Inference - Confidence Intervals

- General (1 - α) Confidence Intervals: a random interval, based on the data, that will include the true population parameter parameter (1 - α) of the time
- ↑confidence → wider interval
- ↑sample size → smaller SE → narrower interval

Confidence Intervals - Example

Suppose gestational times are normally distributed with a standard deviation of 6 days. A sample of 30 mothers yields a mean pregnancy length of 279.5 days.

- What is the "best" estimate of the population mean length of pregnancies?
- Can we provide an interval which is highly likely to include the population mean length of pregnancies?

Confidence Intervals

When we do not know the population parameter, how can we use the sample to estimate the population mean, and use our knowledge of probability to give a range of values consistent with the data?

Parameter: μ

Estimate: $\overline{\mathrm{X}}$

Given a normal population, or large sample size, we can state:

$$P\left[-1.96 \leq \frac{\overline{\mathrm{X}}-\mu}{\sigma/\sqrt{\mathrm{n}}} \leq +1.96\right] = 0.95$$

Note: this is $\frac{\sigma}{\sqrt{n}}$ not $\sigma\sqrt{n}$

Confidence Intervals

$$P\left[-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le +1.96\right] = 0.95$$

… algebra …

$$P\left[\overline{X}-1.96\sigma/\sqrt{n} \le \mu \le \overline{X}+1.96\sigma/\sqrt{n}\right] = 0.95$$

The interval

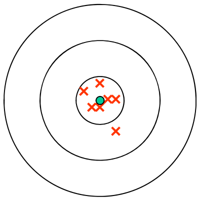
$$\left(\overline{\mathrm{X}}-1.96\sigma/\sqrt{n}\,,\overline{\mathrm{X}}+1.96\sigma/\sqrt{n}\right)$$

is called a **95% confidence interval for** μ.

Interpretation: If we repeat this procedure 100 times, the interval constructed in this manner will include the true mean (μ) 95 times.

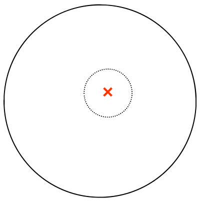
Confidence Intervals – an analogy

Suppose an archer is very accurate – 95% of his shots are within 10 cm of the bulls-eye



Now suppose you are behind the target and can see where the arrow hits but can’t see the bulls-eye. How can you estimate the position of the bulls-eye?

Confidence Intervals – an analogy



Draw a circle of circumference 10 cm around the arrow … you are 95% confident that the circle includes the bulls-eye (i.e. your circle will include the bulls-eye for 95 arrows out of 100).

Confidence Interval: Population Mean

- In general, a 100(1-α)% confidence interval for μ is

$$\left(\overline{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$$

Since the normal distribution is symmetric, we can replace $z_{\alpha/2}$ with $-z_{1-\alpha/2}$ giving …

$$\left(\overline{X} - z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$$

Confidence Intervals

σ known - EXAMPLE

Suppose gestational times are normally distributed with a standard deviation of 6 days. A sample of 30 mothers yields a mean pregnancy length of 279.5 days. Construct a 90% confidence interval for the mean length of pregnancies based on this sample.

Summary

- General (1 - α) Confidence Intervals: (statistic $\pm$ quantile$_{1-\alpha/2}$ × SE$_{\text{statistic}}$)
- CI for μ with σ known → z.
- ↑confidence → wider interval
- ↑sample size → smaller SE → narrower interval

"Without theory (hypotheses), data are meaningless" – William Deming

Statistical Inference - Hypothesis Testing

- Null Hypothesis
- Alternative Hypothesis
- p-value
- Statistically significant
- One-sided (one-tailed) test
- Two-sided (two-tailed) test
- power
- Types of errors: Type I (α), Type II (β)

Hypothesis Testing
Introduction

- In the late 1980's a number of observational studies suggested that infection with various STI increased the risk of acquiring HIV.
- Biologically, it made sense.
- But observational studies are subject to confounding, so …
- A community randomized trial was initiated in Mwanza, Tanzania.
    - Intervention: Enhanced syndromic tx of STI vs SOC
    - 6 paired communities – in each pair, one got the intervention, one got SOC
    - Endpoint: HIV incidence in "sentinel" cohorts
    - Result: Intervention did better in each pair!! (Grosskurth, 1995)

Hypothesis Testing
Introduction

➢ How to interpret? A formal analysis …

- Define
  - p = Prob(intervention community has lower incidence)
  - X = number of pairs where intervention did better
- We want to choose between these two hypotheses
  - Ho: p = 0.5 (i.e. like flipping a coin)
  - Ha: p > 0.5
- If Ho true (intervention is no better than SOC) then probability that *all 6 pairs would favor the intervention* is ...

Hypothesis Testing
Introduction

➢ How to interpret? A formal analysis …

- Define
  - p = Prob(intervention community has lower incidence)
  - X = number of pairs where intervention did better
- We want to choose between these two hypotheses
  - Ho: p = 0.5 (i.e. like flipping a coin)
  - Ha: p > 0.5
- If Ho true (intervention is no better than SOC) then probability that *all 6 pairs would favor the intervention* is ... 1 in 64

Hypothesis Testing Introduction

- How to interpret? A formal analysis …
  - Define

    p = Prob(intervention community has lower incidence)

    X = number of pairs where intervention did better
  - We want to choose between these two hypotheses

    Ho: p = 0.5 (i.e. like flipping a coin)

    Ha: $p > 0.5$
  - If Ho true (intervention is no better than SOC) then probability that *all 6 pairs would favor the intervention* is ... 1 in 64
  - If Ha true (intervention is better than SOC) then probability that *all 6 pairs would favor the intervention* is ... higher
- Need to decide … Ho true or Ha true?

Hypothesis Testing Introduction

- What is the probability of a wrong decision?

	H_0 True	H_A True
Decide H_0	$1-\alpha$	β (type II error)
Decide H_A	α (type I error)	$1-\beta$

α = significance level = P(reject H_0 | H_0 true)

$1-\beta$ = power = P(reject H_0 | H_A true)

- For now, focus on type I error rate (α) only.
- We can choose α *even before collecting any data*.

Hypothesis Testing
Introduction

- Suppose we choose $\alpha = 0.05$.
  - Should I reject Ho if $X = 6$? $P(X = 6 \mid p = .5) = 1/64 < 0.05$ (ok)
  - Should I reject Ho if $X \geq 5$? $P(X \geq 5 \mid p = .5) = .11 > .05$ (not ok)
- Summary
  - Ho: $p = 0.5$ (intervention not effective)
Ha: $p > 0.5$ (intervention effective)
  - Test statistic = X (number of pairs where intervention won)
  - Maximum allowable type I error rate = 0.05
  - Decision rule: Reject if $X = 6$
  - Decision: I observed $X = 6$, therefore reject Ho and conclude STI control does reduce HIV incidence

Hypothesis Testing
Introduction

- What is smallest α I could have chosen and still decided to reject with the observed data ($X = 6$)?
- P-value:
  1. Smallest α that still allows you to reject with the observed data
  2. Probability of data as extreme or more extreme as the observed, assuming Ho is true
- Decision: I observed $X = 6$, therefore reject Ho and conclude STI control does reduce HIV incidence (p-value = .016)
- A happy ending?

Hypothesis Testing
Introduction

- Unfortunately, 4 subsequent trials have failed to replicate the Mwanza result
- Was it …
  - The details of the intervention?
  - Stage of the (HIV) epidemic?
  - Types of STI's?
  - Risk behaviors?
- Or could it have been … chance?

Hypothesis Testing

A second example:

- Among 20-74 year-old males the mean serum cholesterol is 211 mg/ml with standard deviation of 45 mg/ml.
- In a sample of 25 hypertensive men we find a mean serum-cholesterol level of 220 mg/ml.
- Is this strong enough evidence to convince us that the mean for the population of hypertensive men differs from 211 mg/ml?

Hypothesis Testing

Define:

$\mu =$ population mean serum cholesterol for male hypertensives

Hypotheses:

1. Null Hypothesis:

   H_0: $\mu = 211$ mg/ml

2. Alternative Hypothesis:

   H_A: $\mu \neq 211$ mg/ml

- Typically, the alternative hypothesis is the thing you are trying to prove.
- The null hypothesis is formalized skepticism … we will assume the null is true until the evidence against it becomes overwhelming

Hypothesis Testing

- Test Statistic?

Hypothesis Testing

- Test Statistic? $\overline{X}$

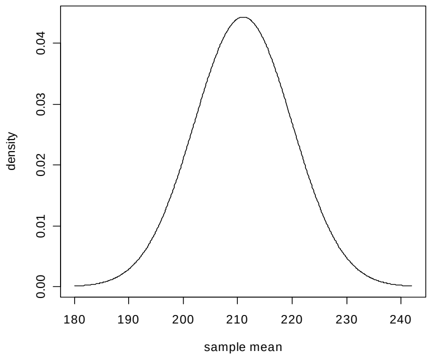
Hypothesis Testing

- Test Statistic? $\overline{X}$

- Distribution of $\overline{X}$? (**assuming Ho true**)

Hypothesis Testing

- Test Statistic? $\overline{X}$
- Distribution of $\overline{X}$? (**assuming Ho true**)

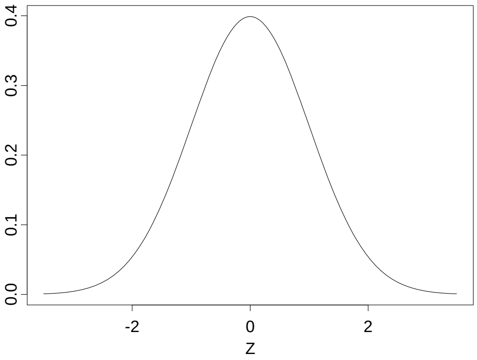


$$\overline{X} \sim N(211, 45^2/25)$$

Hypothesis Testing

$$Z = \frac{\overline{X} - \mu_o}{\sigma/\sqrt{n}}$$

We use μ_o to denote the null value of μ (211 in this case)



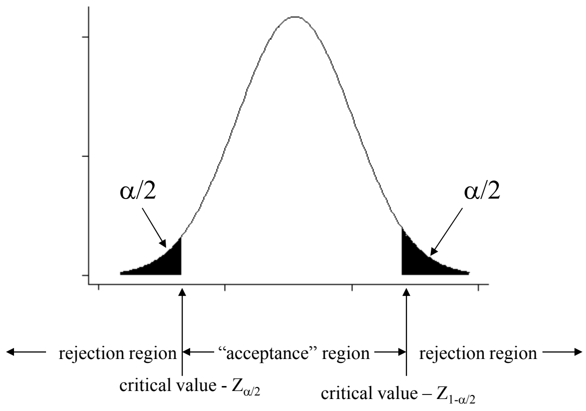
Distribution of Z **if H_0 is true**

Hypothesis Testing

- These plots show us what the distribution of the test statistic ($\overline{X}$) or the standardized test statistic (Z) would look like if we repeated the experiment many times, **and Ho is true**
- **If Ho is true**, we expect values of $\overline{X}$ near 211 or ,equivalently, values of Z near 0.
- Suggests that we should reject Ho if $\overline{X}$ is NOT near 211 (or, equivalently, if Z is NOT near 0) … i.e. reject if we see values that are not likely to occur under Ho assumption
- But how to make sure that our type I error rate is low?

Hypothesis Testing
Critical values, rejection/acceptance regions

- We can choose α before collecting any data. Choosing α tells us what values of Z (or $\overline{X}$) will lead us to reject the null.
- α defines the **critical values** and **rejection regions.**



Hypothesis Testing
Critical values, rejection/acceptance regions

Cholesterol example:

Suppose we choose $\alpha = .05$

Two approaches:

$\Rightarrow$ Critical values on Z-scale are $Z_{\alpha/2} = -1.96$ and $Z_{1-\alpha/2} = 1.96$

$\Rightarrow$ Reject Ho if $Z < -1.96$ or $Z > 1.96$

OR

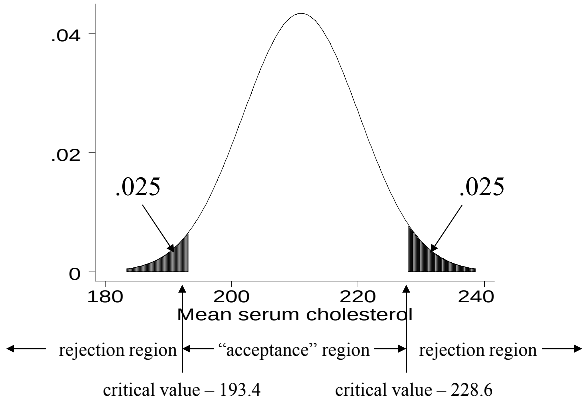
$\Rightarrow$ Critical values on original scale are $\overline{X} = 193.4$ and $\overline{X} = 228.6$

$\Rightarrow$ Reject Ho if $\overline{X} < 193.4$ or $\overline{X} > 228.6$

- Either approach is okay and will give identical answers

Hypothesis Testing
Critical values, rejection/acceptance regions

Cholesterol example (critical values shown in the original scale):



Hypothesis Testing

Cholesterol Example:

We observe $\overline{X}$ = 220 mg/ml or, on the Z scale, assuming $H_0 : \mu = 211$ mg/ml is true,

$$Z=\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}=\frac{220-211}{45/\sqrt{25}}=1.0$$

- Decision – reject Ho or don’t reject? ($\alpha = 0.05$)
- Since Z ($\overline{X}$) is in the “acceptance” region, I decide **not to reject Ho*** (statistical conclusion) and conclude that there is no strong evidence that the mean cholesterol level of hypertensive men is different from 211 (scientific conclusion).
- p-value?

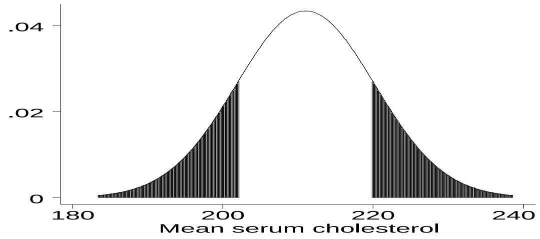
*we say “do not reject Ho” rather than “accept Ho” since failing to disprove a hypothesis is not the same as proving it is true.

Hypothesis Testing - p-value

p-value:

- Smallest α that still allows you to reject with the observed data
- Probability of obtaining data as extreme or more extreme than the actual observed result (given H_0 true).

Cholesterol Example:



Hypothesis Testing - p-value

p-value: Cholesterol Example

$\overline{X} = 220$ mg/ml $n = 25$ $\sigma = 45$ mg/ml

$H_0 : \mu = 211$ mg/ml
$H_A : \mu \neq 211$ mg/ml

The p-value is given by: $P(Z \geq 1) + P(Z \leq -1) =$
$P[\overline{x} > 220] + P[\overline{x} < 202] =$
$2 * P[\overline{x} \geq 220] =$
$2 * P[Z \geq 1.0] = .32$

Rule:
$p < \alpha \Rightarrow$ reject Ho
$p > \alpha \Rightarrow$ do not reject Ho

Guidelines for Judging the Significance of p-value (Rosner pg 219**)**

If $.01 \leq p < .05$, then the results are *significant*.

If $.001 \leq p < .01$, then the results are *highly significant*.

If $p < .001$, then the results are *very highly significant*.

If $p > .05$, then the results are considered *not statistically significant* (sometimes denoted by NS). However, if $.05 \leq p < .10$, then a trend toward statistically significance is sometimes noted.

Statistical significance ≠ scientific significance

Hypothesis Testing
Critical values, rejection/acceptance regions

- The following table summarizes the relationship between α-levels/critical values and p-values/test statistics for a two-tailed test

	Before data	After data	Reject if ..
Probability	α	p	$p < \alpha$
Orig. Test Stat.	C_1, C_2	$\bar{X}$	$\bar{X} < C_1$; $\bar{X} > C_2$
Stand. Test Stat.	$Z_{\alpha/2}, Z_{1-\alpha/2}$	Z	$Z < Z_{\alpha/2}$; $Z > Z_{1-\alpha/2}$

Hypothesis Testing
"how many sides?"

Depending on the alternative hypothesis a test may have a **one-sided alternative** or a **two-sided alternative**. Consider

$$H_0 : \mu = \mu_0$$

We can envision (at least) three possible alternatives

$$H_A : \mu \neq \mu_0 \quad (1)$$

$$H_A : \mu < \mu_0 \quad (2)$$

$$H_A : \mu > \mu_0 \quad (3)$$

(1) is an example of a "two-sided alternative"

(2) and (3) are examples of "one-sided alternatives"

The distinction impacts the p-value calculation and, therefore, our conclusions.

Hypothesis Testing
"how many sides?"

Cholesterol Example: Instead of the two-sided alternative considered earlier we may have only been interested in the alternative that hypertensives had a higher serum cholesterol.

$$H_0 : \mu = 211$$

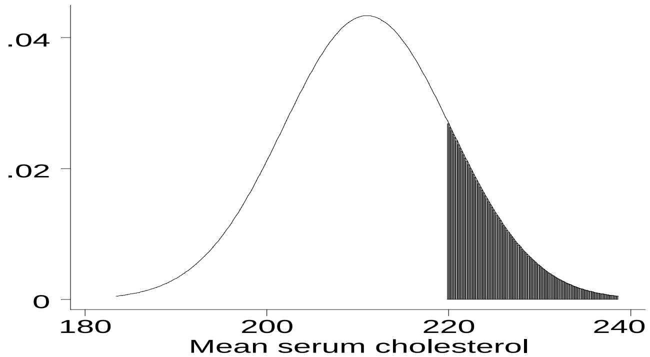
$$H_A : \mu > 211$$

- The p-value would be half of the previous,

$$\begin{aligned} \text{p-value} &= P[\bar{x} > 220 \mid \text{Ho true}] \\ &= P[Z > 1 \mid \text{Ho true}] \\ &= .163 \end{aligned}$$

- In this case, our decision does not change … do not reject Ho.

p-value calculation for Ha: $\mu > 211$ when $\bar{x} = 220$



- sketch the p-value calculation for Ha: $\mu < 211$ if $\bar{x} = 220$

Hypothesis Testing

Through these example we have seen the basic components to the statistical test of a scientific hypothesis.

Summary

1. Identify H_0 and H_A
2. Identify an "appropriate" test statistic
3. Find distribution of the test statistic **assuming Ho true**
4. Compare p-value to the significance (α) level (or test statistic to critical values)
5. Interpret the result

This 5-step procedure is the basis of **all** hypothesis tests, but …

different hypotheses ⇒ different details

Limitations of Hypothesis Testing

What does the p-value tell us?

- If I repeat the experiment many times, what proportion of samples would yield a test statistic further from the null hypothesis than this experiment?
- In other words, it tells us about sampling variability (under Ho assumption)

What doesn't the p-value tell us?

- How far we are from the null
- How big (or small) is the sample
- How variable are the observations

Limitations of Hypothesis Testing

- The p-value also does not tell us the probability that Ho (or Ha) is true or false.
- $P(\text{data} \mid Ho) \neq P(Ho \mid \text{data})$

Limitations of Hypothesis Testing

- If you use multiple tests to evaluate a single scientific hypothesis, the type I error rate will be inflated
- Imagine a clinical trial to test a treatment vs placebo. To ensure that we don’t randomize participants to placebo any longer than necessary, we evaluate the treatment half way through the trial. If the test is significant, we will stop the trial and declare the treatment successful. If not, we continue to the end of the study as planned. What is the overall type I error rate for this trial?

Limitations of Hypothesis Testing

➢ If you use the same data to generate and test a hypothesis, the type I error rate will be inflated.

- Hypothesis testing works properly when the hypothesis is prespecified and data are carefully collected to test the hypothesis.
- If one scans through data looking for “interesting” findings (i.e. secondary or subgroup analyses), the type I error rate will be much higher than anticipated. Such analyses should be thought of as *hypothesis-generating.*

Hypothesis Testing and Confidence Intervals

Confidence Interval: “Plausible” values for μ are given by

$$\overline{X} - z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

If the null hypothesis value of μ (namely, μ_0) is not in this interval, then it is not “plausible” and should be rejected.

In fact, this relationship is “exact” –

1) any value for μ that is not in an α-level confidence interval will be rejected by a two-sided α-level hypothesis test.
2) a 100(1-α)% confidence interval for μ consists of all possible values of μ_0 that would *not* be rejected in a two-sided α-level hypothesis test.

Summary

- •Null Hypothesis
- •Alternative Hypothesis
- •Statistically significant
- •p-value
- •power
- •Types of errors: Type I (α), Type II (β)
- •One-sided (one-tailed) test
- •Two-sided (two-tailed) test
- •Critical value
- •Acceptance/rejection region

Statistical Inference: t-test

t-distribution

To make inferences about μ using the methods outlined previously, we need to know σ^2. But usually, **σ is unknown** - we only have s^2. It turns out that even though

$$\frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}}$$

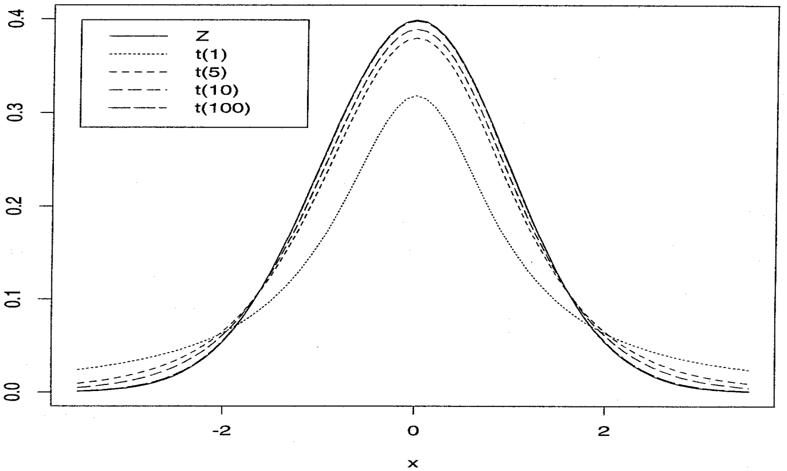
is normally distributed,

$$\frac{(\overline{X}-\mu)}{s/\sqrt{n}}$$

is not (quite)!

W.S. Gosset worked for Guinness Brewing in Dublin, IR. He was forced to publish under the pseudonym "Student". In 1908 he derived the distribution of $\frac{(\overline{X}-\mu)}{s/\sqrt{n}}$ which is now known as Student's **t-distribution**.

Normal and t distributions



Confidence Intervals - σ^2 unknown

When σ is unknown we replace it with the estimate, s, and use the t-distribution. The statistic

$$\frac{\overline{X}-\mu}{s/\sqrt{n}}$$

has a t-distribution with n-1 *degrees of freedom*.

We can use this distribution to obtain a confidence interval for μ and test hypotheses even when σ is not known.

A (1-α)% **Confidence Interval for the Population Mean when σ is unknown is**

$$\left(\overline{X}-t_{n-1,1-\alpha/2}\times s/\sqrt{n},\ \overline{X}+t_{n-1,1-\alpha/2}\times s/\sqrt{n}\right)$$

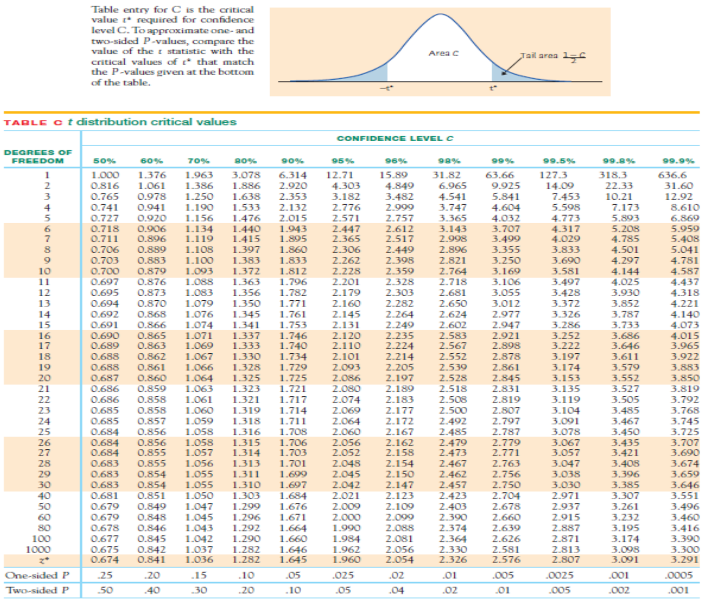


Table entry for C is the critical value t^* required for confidence level C. To approximate one- and two-sided P-values, compare the value of the t statistic with the critical values of t^* that match the P-values given at the bottom of the table.

TABLE C *t* distribution critical values

DEGREES OF FREEDOM	CONFIDENCE LEVEL *C*											
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
One-sided *P*	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
Two-sided *P*	.50	.40	.30	.20	.10	.05	.04	.02	.01	.005	.002	.001

Confidence Intervals - σ^2 unknown t Distribution - EXAMPLE

Given our 30 moms with a mean gestation of 279.5 days and a sample variance of 28.3, compute a 90% confidence interval for the mean length of pregnancies:

Hypothesis Testing 1-sample Tests: t Test

We can also use the **t-distribution** for hypothesis tests.

Going back to the cholesterol example, suppose that in our sample of 25 hypertensive men we observe a sample mean of 220 mg/ml and a sample standard deviation of 38.6 mg/ml and we want to test the hypothesis

Ho: $\mu = 211$

Ha: $\mu \neq 211$

Hypothesis Testing
1-sample Tests: t-test

Cholesterol Example:

$\overline{X} = 220$ mg/ml $s = 38.6$ mg/ml $n = 25$

H_0 : $\mu = 211$ mg/ml

H_A : $\mu \neq 211$ mg/ml

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

$$= \frac{220 - 211}{38.6/\sqrt{25}} = 1.17$$

For an $\alpha = 0.05$ test we use the critical value determined from the t(24) distribution: $t_{24,\ .975} = 2.064$

Since $|T| = 1.17 < 2.064$ the difference is not statistically significant at the $\alpha = 0.05$ level and we fail to reject H_0.

Hypothesis Testing
1-sample Tests: t Test

```
. ttesti 25 220 38.6 211

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      25         220        7.72        38.6    204.0667    235.9333
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   1.1658
Ho: mean = 211                                   degrees of freedom =       24

   Ha: mean < 211              Ha: mean != 211              Ha: mean > 211
 Pr(T < t) = 0.8724         Pr(|T| > |t|) = 0.2551          Pr(T > t) = 0.1276
```

Hypothesis Testing
1-sample Tests

- Z-Test: Test of mean with known variance
- t-Test: Test of mean with unknown variance

Hypothesis Testing
2-sample Test

- In all the previous examples we had a sample of data from a single population and asked if the population mean was equal to a fixed number:

$$Ho: \mu = \mu_0$$

- Perhaps more common is the situation in which you have samples from two different populations (drug and placebo, men and women, etc) and want to know if the population means are equal:

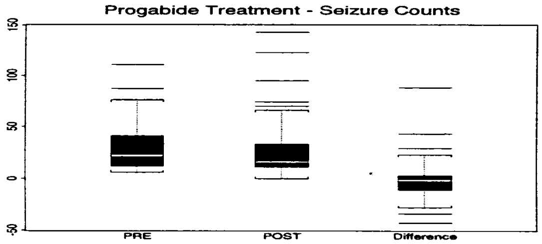
$$Ho: \mu_1 = \mu_2$$

- We will examine two distinct study designs that lead to a two sample test: paired (matched) and unpaired designs

Hypothesis Testing
2-sample Motivation (paired)

Epileptic patients are measured for 4 weeks to record their baseline number of seizures. Patients are then administered treatment with progabide and the number of seizures in the next 4 weeks is recorded. (N = 59 patients).

	Baseline	Treatment	Difference
Mean	29.15	28.41	-0.74
Std Dev	21.89	28.70	19.47



Hypothesis Testing
2-sample: Paired t Test

Assume that we have 2 observations on each person (unit)

X_{1i} = the first observation on person i
X_{2i} = the second observation on person i

The data are naturally paired. The hypothesis of interest is

Ho: $\mu_1 = \mu_2$ or, equivalently, Ho: $\mu_1 - \mu_2 = \mu_d = 0$

We can base inferences on the differences:

$$d_i = X_{1i} - X_{2i}$$

The hypothesis of "no treatment effect" can be addressed with the statistics:

$\bar{d} =$ average change in measurement from time 1 to time 2
$s_d^2 =$ sample variance for the differences

Now we can just apply the one-sample procedure we have learned …

Hypothesis Testing
2-sample: Paired t Test

Hypotheses: $H_0 : \mu_d = 0$

$H_A : \mu_d \neq 0$

The testing of the hypotheses is then based on the assumed distribution of d_i under the null hypothesis. Either we assume the d_i are normally distributed …

$$d_i \sim N\left(0, \sigma_d^2\right)$$

or we appeal to the CLT. Under the null hypothesis,

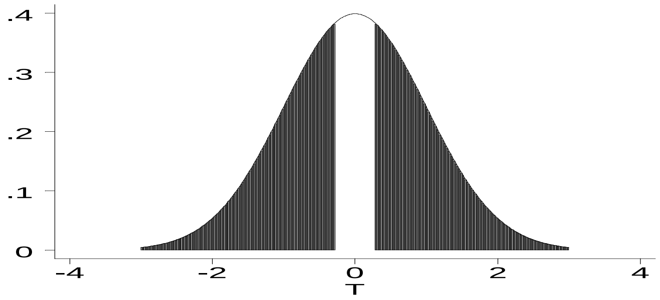
$$T = \frac{\bar{d} - 0}{s_d / \sqrt{n}}$$

has a t-distribution with n - 1 degrees of freedom.

Hypothesis Testing
2-sample: Paired t Test

```
. ttesti 59 -.74 19.47 0

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      59        -.74    2.534778       19.47    -5.81391     4.33391
------------------------------------------------------------------------------
    mean = mean(x)                                                t =  -0.2919
Ho: mean = 0                                     degrees of freedom =       58

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.3857         Pr(|T| > |t|) = 0.7714          Pr(T > t) = 0.6143
```

A sketch of the p-value calculation is given below:



How about a 90% confidence interval for μ_d?

Hypothesis Testing
2-sample Motivation (unpaired)

Suppose that we measure serum iron levels for two groups of children: one group healthy; and one group suffering from cystic fibrosis. We obtain the following data (serum iron measured in μmol/l):

	N	Mean	Std. Dev.
Healthy	9	18.9	5.9
Cystic Fibrosis	13	11.9	6.3

Q: Is the mean serum iron level the same for both populations?

Q: What is a confidence interval for the mean difference?

Hypothesis Testing
2 Independent Samples

We make the following assumptions (justify?)

1. The two samples are independent.
2. The measurements are distributed normally within each population.
3. (or) The sample sizes are “large”.

What are the parameters and statistics in this situation?

		Group 1	Group 2
Population	Mean	μ_1	μ_2
	Std Dev	σ_1	σ_2
Sample	Mean	$\overline{X}_1$	$\overline{X}_2$
	Std Dev	s_1	s_2
	sample size	n_1	n_2

Hypothesis Testing
2 Independent Samples

As before the hypotheses often are:

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

Q: What statistic addresses this hypothesis?

$$\overline{X}_1 - \overline{X}_2$$

Q: What is the distribution of this statistic?

$$\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

This "simply" comes from two facts that we've seen before:

1. sums of **independent** normals are **normal** with mean equal to the **sum of the component means** and variance equal to the **sum of the component variances**.

2. The distributions of the sample means.

$$\overline{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$$
$$\overline{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

(Aside: why doesn't this argument work in the paired case?)

Hypothesis Testing
2 Independent Samples

For testing the hypotheses

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

Use the **standardized** test statistic (assuming σ_1 and σ_2 unknown)

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Q: What is the distribution of this statistic?
A: It's approximately t-distributed. To use the t-distribution we need the "degrees of freedom". We can use the smaller of n_1-1 and n_2-1, or, preferably, Stata computes the approximate degrees of freedom (complex formula).

Hypothesis Testing
2 Independent Samples - Unknown Variances

```
. ttesti 9 18.9 5.9 13 11.9 6.3, unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |       9        18.9    1.966667         5.9    14.36486    23.43514
       y |      13        11.9    1.747306         6.3    8.092948    15.70705
---------+--------------------------------------------------------------------
combined |      22    14.76364    1.482474     6.95342    11.68066    17.84661
---------+--------------------------------------------------------------------
    diff |                   7    2.630752                 1.47512    12.52488
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)                                      t =   2.6608
Ho: diff = 0                     Satterthwaite's degrees of freedom =  18.0971

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9921         Pr(|T| > |t|) = 0.0159          Pr(T > t) = 0.0079
```

Comments:

1. For our purposes, always use "unequal" option
2. Note the 3 p-values; choose the one corresponding to your alternative hypothesis

Confidence Interval - 2 samples

We can form a confidence interval for the mean difference $\mu_d = \mu_1 - \mu_2$. In general, use

$$\left((\bar{X}_1 - \bar{X}_2) - t_{df,1-\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},\ (\bar{X}_1 - \bar{X}_2) + t_{df,1-\alpha/2} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

where df is the appropriate degrees of freedom.

Summary

1. Testing for $\mu_1=\mu_2$.
2. Issues:
    a. paired vs unpaired (independent)
    b. variances known vs not known

Statistical Inference - Proportions

One sample
- Confidence intervals
- Hypothesis tests

Two Sample
- Confidence intervals
- Hypothesis tests

Hypothesis Testing
1-sample Tests: Binomial Proportion

Example:

Suppose that there is an equal chance that a child is male or female. In a sample of 114 workers at a pesticide plant (with only one child) we find that 67 of the children are female. Is this evidence that these workers are more likely to have girls? What is a confidence interval for the proportion of female children?

Define: p = probability that worker has daughter

$$H_0 : p = 0.5$$

$$H_A : p > 0.5$$

Hypothesis Testing
1-sample Tests: Binomial Proportion

Recall the definition of a p-value …

p-value = Pr (as extreme or more extreme | Ho true)

= Pr($X \geq 67 \mid n = 114, p = .5$)

= $P(X=67|n=114,p=.5) + P(X=68|n=114,p=.5) + \ldots + P(X=114|n=114,p=.5)$

where

$$P(X = k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Valid … but looks like a lot of work …

Confidence Intervals -
Binomial Proportion

Recall that if n is “large” ($np \geq 10$ and $n(1-p) \geq 10$) we can approximate the binomial distribution by a normal:

$$X \sim N(np, np(1-p))$$

or, equivalently,

$$\hat{p} = \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

We can use the normal appoximation to test hypotheses and generate confidence intervals about p …

Hypothesis Testing
1-sample Tests: Binomial Proportion

Data: $X = 67$, $n = 114$

$$X \sim N(\,np,\ np(1-p)\,)$$

Under H_0: $p = 0.5$

$$Z = \frac{X - n \times p_0}{\sqrt{n \times p_0 \times (1 - p_0)}}$$

$$= \frac{67 - 114 \times 0.5}{\sqrt{114 \times 0.5 \times (1 - 0.5)}} = 1.87$$

For the one-sided alternative H_a: $p > 0.5$ we reject H_0 if $Z > z_{1-\alpha}$

The critical value for a one-sided $\alpha = 0.05$ test is $z_{1-.05} = 1.65$. Since the test statistic, $Z = 1.87$, exceeds the critical value, we reject H_0.

Repeat this using $\hat{p}$ in place of X to convince yourself it is equivalent!

Confidence interval? (use $\hat{p}$ instead of p_o in the standard error)

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Hypothesis Testing
1-sample Tests: Binomial Proportion

```
. prtesti 114 67 .5, count

One-sample test of proportion                   x: Number of obs =      114
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.                     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .5877193    .046103                      .4973591    .6780795
------------------------------------------------------------------------------
    p = proportion(x)                                             z =   1.8732
Ho: p = 0.5

     Ha: p < 0.5                 Ha: p != 0.5                 Ha: p > 0.5
 Pr(Z < z) = 0.9695         Pr(|Z| > |z|) = 0.0610          Pr(Z > z) = 0.0305
```

```
. bitesti 114 67 .5

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
      114         67           57       0.50000      0.58772

  Pr(k >= 67)            = 0.037346  (one-sided test)
  Pr(k <= 67)            = 0.975638  (one-sided test)
  Pr(k <= 47 or k >= 67) = 0.074693  (two-sided test)
```

Hypothesis Testing
2 Sample Proportions - Motivation

In a study of the morbidity and mortality among pediatric victims of motor vehicle accidents, information was gathered on whether children were wearing seat belts at the time of the accident (Osberg and DiScala, AJPH 1992). Of the 123 children who were wearing a seat belt, 3 died, while of the 290 children found not wearing a seat belt, 13 died.

Q: Does this study show that wearing a seat belt is beneficial?

Let’s construct a hypothesis test to assess the two fatality probabilities.

Hypothesis Testing
2 Sample Proportions

What are the parameters and the statistics for this question?

	Group 1	Group 2
Population Proportion	p_1	p_2
Sample Proportion	$\hat{p}_1=X_1/n_1$	$\hat{p}_2=X_2/n_2$
Sample Size	n_1	n_2

p_1 = mortality for children when seat belt is worn
p_2 = mortality for children when seat belt is not worn
X_1 = # of recorded deaths among those who wear seat belts
n_1 = # of children surveyed from those that wear seat belts
X_2 = # of recorded deaths among those who do not wear seat belts
n_2 = # of children surveyed from those that no not wear seat belts

Hypothesis Testing
2 Sample Proportions

The hypothesis that the two populations are the same is addressed by the hypotheses:

$$H_0 : p_1 = p_2$$
$$H_A : p_1 \neq p_2$$

A statistic useful for this comparison is the difference in the observed, or sample, proportions (we'll see some others later):

$$\hat{p}_1 - \hat{p}_2$$

Q: What is the distribution of this statistic?

A: Approximately normal.

$$\hat{p}_1 - \hat{p}_2 \sim N\left(\mu = p_1 - p_2,\ \sigma^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

Hypothesis Testing
2 Sample Proportions

	Group	Fatalities	Sample Size
1	seat belts +	3	123
2	seat belts -	13	290

So we estimate the separate risks as

$$\hat{p}_1 = 3/123 = 0.024$$
$$\hat{p}_2 = 13/290 = 0.045$$

We can use Stata to test

$$H_0: p_1 = p_2 \, (= p_0)$$

(What is the estimate of the common fatality rate, p_0, under the null?)

Hypothesis Testing
2 Sample Proportions

```
. prtesti 123 .024 290 .045

Two-sample test of proportion                     x: Number of obs =      123
                                                  y: Number of obs =      290
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |       .024      .0138                      -.0030474    .0510474
           y |       .045   .0121733                       .0211407    .0688593
-------------+----------------------------------------------------------------
        diff |      -.021   .0184019                       -.057067     .015067
             |  under Ho:   .0207661    -1.01   0.312
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                  z =  -1.0113
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.1559         Pr(|Z| < |z|) = 0.3119          Pr(Z > z) = 0.8441
```

Conclusion?

Why two different std errors?

Hypothesis Testing
2 Sample Proportions

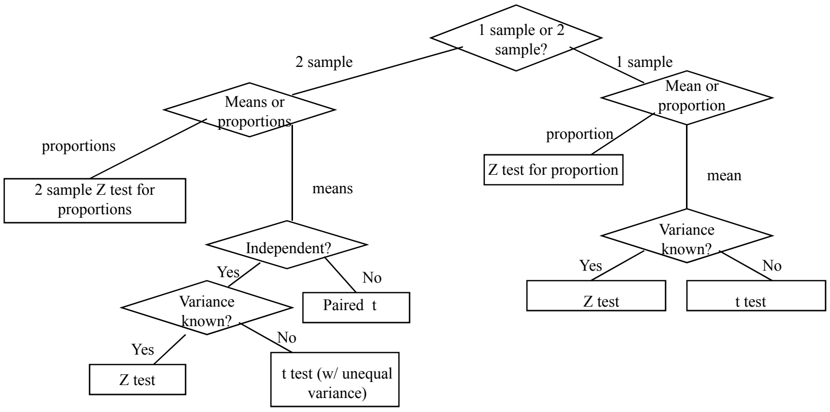
Note, under the null we would have expected:

n_1 x p_0 = 0.039 X 123 = 4.73 deaths in the seat belt + group.

n_2 x p_0 = 0.039 X 290 = 11.15 deaths in the seat belt - group.

The fact that one of these is “small” (less than 5) causes some concern about the normal approximation to the binomial. One alternative in this case is known as **Fisher’s Exact Test** - which does not make the normality assumption.

Flow chart for hypothesis testing



1 sample or 2 sample?
2 sample
1 sample
Means or proportions
proportions
2 sample Z test for proportions
means
Independent?
Yes
No
Paired t
Variance known?
Yes
Z test
No
t test (w/ unequal variance)
Mean or proportion
proportion
Z test for proportion
mean
Variance known?
Yes
Z test
No
t test

Remember: Always verify assumptions before using a test!

Power and Sample Size

- Power = Pr(reject H_0 | H_a true)
- Power depends on μ_0, μ_a, σ^2, and n.
- Sample size calculations ensure that the study is capable of detecting departures from the null hypothesis.
- Power and Sample size require a model for the data under **both** the null and the alternative is required.

Power and Sample Size

We have mainly focused on the distribution of the test statistic under the null hypothesis. Shouldn't we also consider the distribution under the alternative hypothesis?

	H_0 True	H_A True
Decide H_0	1-α	β (type II error)
Decide H_A	α (type I error)	1-β

α = P(reject H_0 | H_0 true)

power (1-β) = P(reject H_0 | H_A true)

Power and Sample Size

- The distribution of the test statistic under the alternative hypothesis tells us the **power** of the test.
- Power indicates the ability of the test procedure to reliably detect departures from the null hypothesis.
- Power ($1 - \beta$) and significance (α) are important considerations in the planning of a study.
- Sample size calculations depend on power (and visa-versa).

Power

Power refers to the probability of rejecting the null hypothesis when it truly is false:

$$\text{power} = P[\text{ reject } H_0 \mid H_A \text{ true }]$$

So when we consider power we compute probabilities assuming that the alternative is the "truth".

Consider the 1-sample testing situation with the hypotheses:

$$H_0 : \mu = \mu_0$$
$$H_A : \mu \neq \mu_0$$

The scenario is enough to illustrate all of the important concepts. Details change when we consider variants:

1. One-sided alternatives
2. 2-sample problems
3. Sample proportions

Power

Power refers to the likelihood of detecting a difference from H_0.

Define $\Delta = | \mu_0 - \mu_A |$

where μ_0 and μ_A are the null and alternative means, respectively.

- ➢ Δ "big" ⇒ easy to detect a difference
- ➢ Δ "small" ⇒ hard to detect a difference
- ➢ Measure "big" and "small" in terms of standard error of the mean units i.e. Δ/sem where sem = $\sigma/\sqrt{n}$

Then what we want "large" to give good power is $\frac{\sqrt{n}\Delta}{\sigma}$

We expect **power to increase** as:

- Sample size increases.
- Distance between μ_0 and μ_A (Δ) increases.
- Variance gets smaller.

Power

Q: How can we compute the power?

We assume that σ is known. Even if we don't know it, we'll need an estimate of it. Also, we know the sample size, n.

1. Choose α and 1-sided / 2-sided

2. Identify μ_0 (usually obvious)

3. Identify μ_A (many possible values; how to choose?)

(1) & (2) determine what the rejection region will be. For example, a two-sided test of H_0: $\mu=\mu_0$ gives the rejection region

REJECT H_0: $\overline{X} > \mu_0 + z_{1-\alpha/2}\sigma/\sqrt{n}$

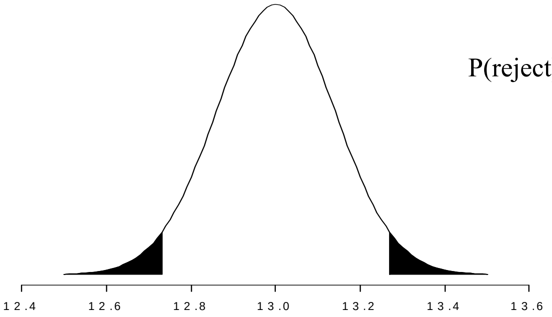
or $\overline{X} < \mu_0 - z_{1-\alpha/2}\sigma/\sqrt{n}$

Power

Ho: μ = 13

Ha: μ ≠ 13

Assume **μ = 13.0**, σ= 0.7, n = 25, α=.05 (Ho true)



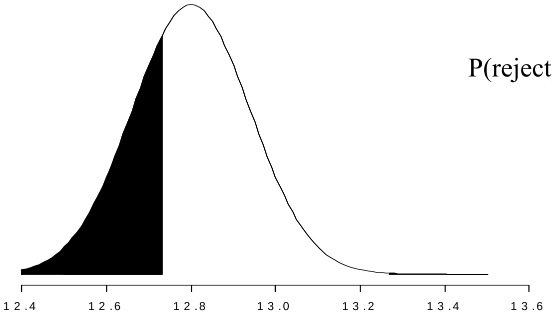
P(reject) = .05

Power

Ho: μ = 13

Ha: μ ≠ 13

Assume **μ = 12.8**, σ= 0.7, n = 25, α=.05   (Ha true)



P(reject) = .31

Power

Notice that for $\mu_A < \mu_0$, only the lower rejection region has non-negligible probability. If $\mu_A > \mu_0$, then only the upper region matters. In any event, we find

$$\begin{aligned}\text{power} &= \mathrm{P}\left[\text{REJECT } \mathrm{H}_0 \mid \mathrm{H}_A \text{ true}\right] \\ &= \vdots \\ &= \text{lots of algebra} \\ &= \vdots \\ &= \mathrm{P}\left(Z < \frac{\sqrt{n}\Delta}{\sigma} - z_{1-\alpha/2} \right)\end{aligned}$$

use $z_{1-\alpha}$ for 1-tailed test

Given n, α, Δ, σ we can look up the power in normal table.

Power

POWER can be written as

$$\mathrm{P}\left(Z < \frac{\sqrt{n}\Delta}{\sigma} - z_{1-\alpha/2} \right)$$

This tells us the **probability of rejecting** H_0 **when the alternative is true**. This is important!! Why spend $$$ on a study, that hopes to show a treatment effect, if the probability of rejecting H_0 is small? In fact, in the acquisition of $$$ from NIH one must show that the study is capable (sufficient power) of detecting a meaningful difference.

Power - Example

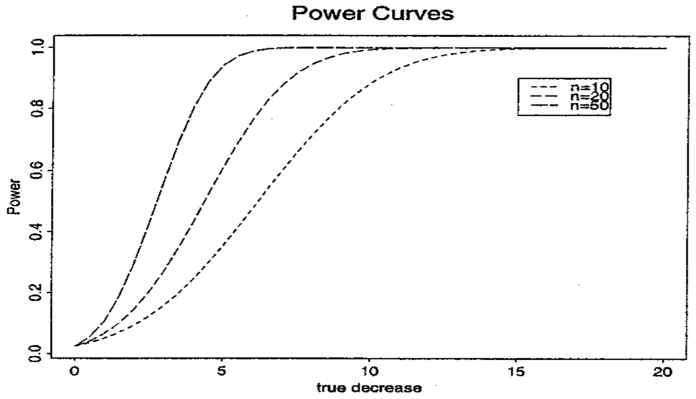
Suppose we are testing blood pressure medication among hypertensive men. We want to design a study that can detect a clinically relevant decline in blood pressure due to the treatment. Suppose 5 mmHg is considered clinically relevant and I enroll 50 participants in the study. Suppose the standard deviation of changes is 10. What is the power ($\alpha = 0.05$, 2-sided)?

$$H_0 : \mu = 0$$
$$H_A : \mu = 5$$

```
. sampsi 0 5, onesample n(50) sd(10) alpha(.05)

Estimated power for one-sample comparison of mean to hypothesized value

Test Ho: m = 0, where m is the mean in the population

Assumptions:
         alpha =   0.0500  (two-sided)
 alternative m =        5
            sd =       10
 sample size n =       50

Estimated power:
         power =   0. 9424
```

Power Curve

One useful display is the **Power Curve**. This shows the power for various μ_A.



Sample Size

Q: How many patients should we treat?

1. Specify significance level (α)
2. Specify null mean (μ_0)
3. Specify power ($1 - \beta$)
4. Specify the alternative mean (μ_A)
5. Specify the variance (σ)

$$n = \frac{\left(z_{1-\beta} + z_{1-\alpha/2}\right)^2 \sigma^2}{\Delta^2}$$

Typical values ….

$z_{.975} = 1.96$ ($\alpha = .05$, 2-sided test)

$z_{.90} = 1.28$ (90% power)

$z_{.80} = 0.84$ (80% power)

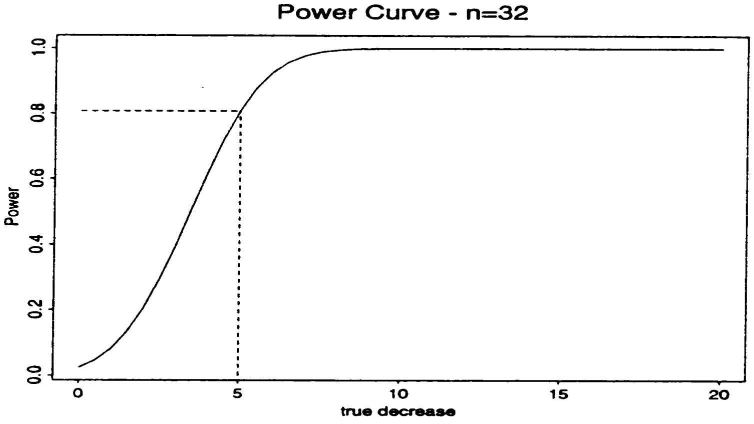
Sample Size - Example

Same situation as before. How many patients are required to obtain 80% power using a 2-sided $\alpha = 0.05$ test?

```
. sampsi 0 5, onesample power(.80) sd(10) alpha(.05)

Estimated sample size for one-sample comparison of mean to hypothesized value

Test Ho: m = 0, where m is the mean in the population

Assumptions:
         alpha =   0.0500  (two-sided)
         power =   0.8000
 alternative m =        5
            sd =       10

Estimated required sample size:
             n =       32
```

- Recruit n = 32 patients into our study.

Sample Size



Factors that Influence Sample Size

The required sample size **increases** as:

1. σ^2 increases
2. significance level is made smaller (α decreases)
3. power increases ($1 - \beta$ increases)
4. distance $| \mu_0 - \mu_A |$ decreases

Sample Size

1-sample Mean, 1-sided Test

$$n = \frac{\sigma^2\left(z_{1-\beta} + z_{1-\alpha}\right)^2}{\left(\mu_0 - \mu_A\right)^2}$$

1-sample Proportion, 1-sided Test

$$n = \frac{p_0\left(1-p_0\right)\left(z_{1-\alpha} + z_{1-\beta}\sqrt{\frac{p_A\left(1-p_A\right)}{p_0\left(1-p_0\right)}}\right)^2}{\left(p_0 - p_A\right)^2}$$

1-sample Proportion, 2-sided Test

$$n = \frac{p_0\left(1-p_0\right)\left(z_{1-\alpha/2} + z_{1-\beta}\sqrt{\frac{p_A\left(1-p_A\right)}{p_0\left(1-p_0\right)}}\right)^2}{\left(p_0 - p_A\right)^2}$$

Sample Size

2-sample Mean, 2-sided Test, Equal sizes ($n_1 = n_2 = n$):

$$n = \frac{\left(\sigma_1^2 + \sigma_2^2\right)\left(z_{1-\beta} + z_{1-\alpha/2}\right)^2}{\left(\mu_0 - \mu_A\right)^2}$$

2-sample Proportion, 2-sided Test, Equal sizes ($n_1 = n_2 = n$):
(rough approximation - for $p_1 \approx p_2$)

$$n = \frac{[p_1\left(1-p_1\right) + p_2\left(1-p_2\right)]\left(z_{1-\beta} + z_{1-\alpha/2}\right)^2}{\left(p_1 - p_2\right)^2}$$

Summary

- Power is an important component in study design.
- Sample size calculations ensure that the study is capable of detecting departures from the null hypothesis.
- Power and Sample size require more than test - a model for the data under **both** the null and the alternative is required.
- STRUTs (α=.05; power = .80; 2-tailed)

  *One sample: $n = 8/D^2$

  *Two sample: $n = 16/D^2$

  where $D = \Delta/\sigma$ (difference in std. dev. units)