UCLA STAT 10

Introduction to Statistical Reasoning

●**Instructor**: **Ivo Dinov,**

Asst. Prof. In Statistics and Neurology

●**Teaching Assistants:** **Yan Xiong, Will Anderson**

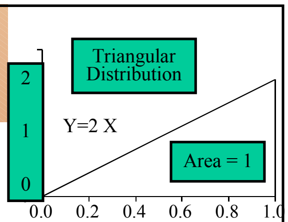
University of California, Los Angeles, Winter 2002

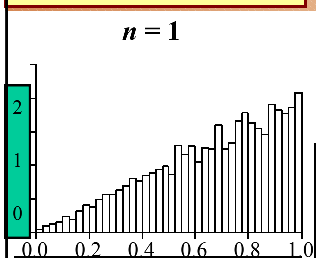
http://www.stat.ucla.edu/~dinov/

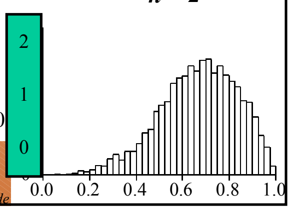
Recall we looked at the sampling distribution of $\overline{X}$

- For the sample mean calculated from a random sample, $E(\overline{X}) = \mu$ and $SD(\overline{X}) = \sigma/\sqrt{n}$, provided $\overline{X} = (X_1+X_2+ \ldots + X_n)/n$, and $X_k \sim N(\mu, \sigma)$. Then
- $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. And variability from sample to sample in the ***sample-means*** is given by the variability of the individual observations divided by the square root of the sample-size. In a way, averaging decreases variability.

Central Limit Effect – Histograms of sample means



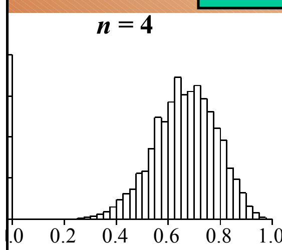


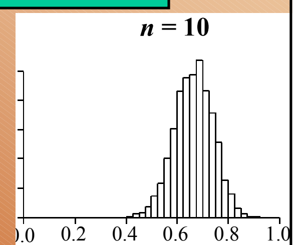


Sample means from sample size n=1, n=2, 500 samples

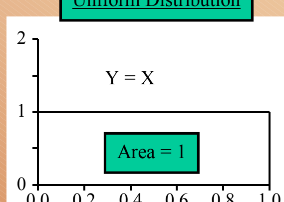
Central Limit Effect -- Histograms of sample means

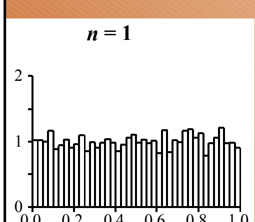
Triangular Distribution Sample sizes n=4, n=10

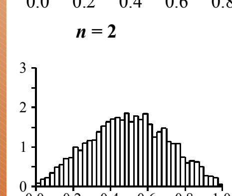




Central Limit Effect – Histograms of sample means



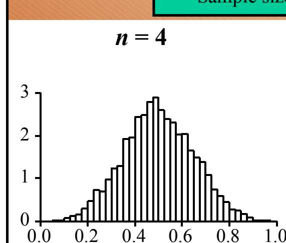


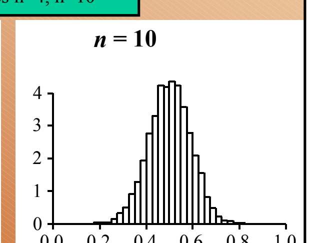


Sample means from sample size n=1, n=2, 500 samples

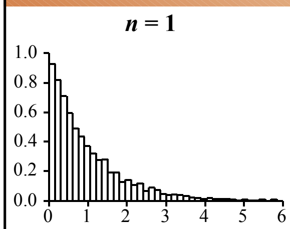
Central Limit Effect -- Histograms of sample means

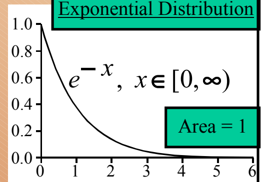
Uniform Distribution Sample sizes n=4, n=10

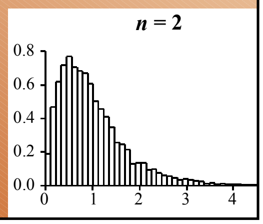




Central Limit Effect – Histograms of sample means



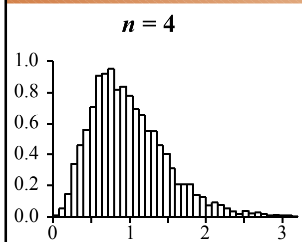


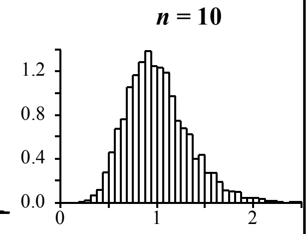


Sample means from sample size n=1, n=2, 500 samples

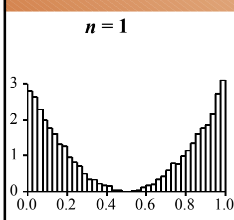
Central Limit Effect -- Histograms of sample means

Exponential Distribution
Sample sizes n=4, n=10

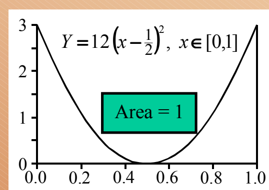


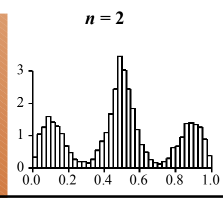


Central Limit Effect – Histograms of sample means



Quadratic U Distribution

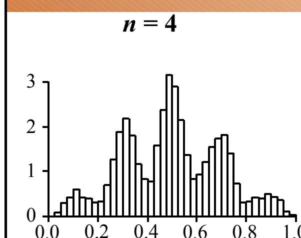


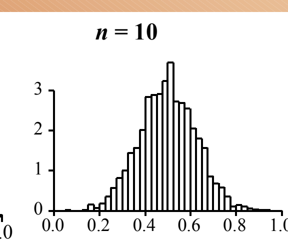


Sample means from sample size n=1, n=2, 500 samples

Central Limit Effect -- Histograms of sample means

Quadratic U Distribution
Sample sizes n=4, n=10





Central Limit Theorem – heuristic formulation

Central Limit Theorem:
When sampling from almost any distribution,
$\overline{X}$ is approximately Normally distributed in large samples.

Show Sampling Distribution Simulation Applet:
file:///C:/Ivo.dir/UCLA_Classes/Winter2002/AdditionalInstructorAids/SamplingDistributionApplet.html

Central Limit Theorem – theoretical formulation

Let $\{X_1, X_2, \ldots, X_k, \ldots\}$ be a sequence of independent observations from one specific random process. Let and $E(X) = \mu$ and $SD(X) = \sigma$ and both be finite ($0 < \sigma < \infty$; $|\mu| < \infty$). If $\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$, sample-avg,

Then $\overline{X}$ has a distribution which approaches $N(\mu, \sigma^2/n)$, as $n \to \infty$.

Review

- What does the central limit theorem say? Why is it useful? (If the sample sizes are large, the mean in Normally distributed, as a RV)
- In what way might you expect the central limit effect to differ between samples from a symmetric distribution and samples from a very skewed distribution? (Larger samples for non-symmetric distributions to see CLT effects)
- What other important factor, apart from skewness, slows down the action of the central limit effect?

(Heavyness in the tails of the original distribution.)

Review

- When you have data from a moderate to small sample and want to use a normal approximation to the distribution of $\overline{X}$ in a calculation, what would you want to do before having any faith in the results? (30 or more for the sample-size, depending on the skewness of the distribution of X. Plot the data - non-symmetry and heavyness in the tails slows down the CLT effects).
- Take-home message: CLT is an application of statistics of paramount importance. Often, we are not sure of the distribution of an observable process. However, the CLT gives us a theoretical description of the distribution of the sample means as the sample-size increases ($N(\mu, \sigma^2/n)$).

The standard error of the mean – remember …

- For the sample mean calculated from a random sample, SD($\overline{X}$) = $\frac{\sigma}{\sqrt{n}}$. This implies that the variability from sample to sample in the *sample-means* is given by the variability of the individual observations divided by the square root of the sample-size. In a way, averaging decreases variability.
- Recall that for *known* SD(X)=σ, we can express the SD($\overline{X}$) = $\frac{\sigma}{\sqrt{n}}$. How about if SD(X) is *unknown*?!?

The standard error of the mean

The standard error of the sample mean is an estimate of the *SD* of the sample mean

- i.e. a measure of the precision of the sample mean as an estimate of the population mean
- given by $\text{SE}(\bar{x}) = \frac{\text{Sample standard deviation}}{\sqrt{\text{Sample size}}}$

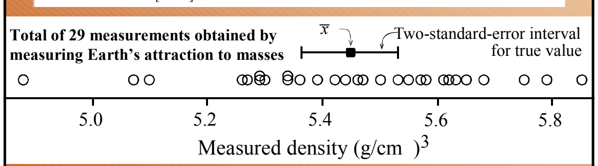
$$SE(\bar{x}) = \frac{s_x}{\sqrt{n}}$$

- Note similarity with
- SD($\overline{X}$) = $\frac{\sigma}{\sqrt{n}}$.

Cavendish's 1798 data on mean density of the Earth, g/cm³, relative to that of H_2O

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Source: Cavendish [1798].



Newton's law of gravitation: $F = G\, m_1\, m_2 / r^2$, the attraction force F is the ratio of the product (Gravitational const, mass of body1, mass body2) and the distance between them, r. Goal is to estimate G!

Cavendish's 1798 data on mean density of the Earth, g/cm³, relative to that of H_2O

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Source: Cavendish [1798].

Sample mean $\bar{x} = 5.447931 \; g/cm^3$

and sample SD = $s_X = 0.2209457 \; g/cm^3$

Then the standard error for these data is:

$$SE(\overline{X}) = \frac{s_X}{\sqrt{n}} = \frac{0.2209457}{\sqrt{29}} = 0.04102858$$

Cavendish's 1798 data on mean density of the Earth, g/cm^3, relative to that of H_2O

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Source: Cavendish [1798].

Safely can assume the true mean density of the Earth is within 2 SE's of the sample mean!

$$\bar{x} \pm 2\times SE(\bar{x}) = 5.447931 \pm 2\times 0.04102858 g/cm^3$$

Review

- Why is the standard deviation of $\bar{X}$, SD($\bar{X}$), not a useful measure of the precision of $\bar{X}$ as an estimator in practical applications? (SD( ) = $\frac{\sigma}{\sqrt{n}}$ and σ is unknown most time!)
- What measure of precision do we use in practice? (SE)
- How is SE($\bar{x}$) related to SD($\bar{X}$)?
- When we use the formula SE($\bar{x}$) = $s_X/\sqrt{n}$, what is s_X and how do you obtain it? (Sample SD(X))

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Review

- What can we say about the true value of μ and the interval $\bar{x} \pm 2$ SE($\bar{x}$) ? (95% sure)
- Increasing the precision of $\bar{x}$ as an estimate of μ is equivalent to doing what to se($\bar{x}$)? (decreasing)

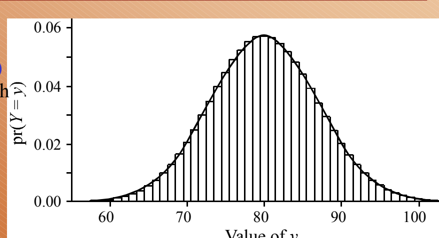
Sampling distribution of the sample proportion

The sample proportion $\hat{p}$ estimates the population proportion ***p***.

Suppose, we poll college athletes to see what percentage are using performance inducing drugs. If 25% admit to using such drugs (in a single poll) can we trust the results? What is the variability of this proportion measure (over multiple surveys)? Could Football, Water Polo, Skiing and Chess players have the same drug usage rates?

Approximate Normality in large samples

Histogram of Bin (200, p=0.4) probabilities with superimposed Normal curve approximation. Recall that for Y~Bin(n,p)



$\mu_Y = E(Y) = np$

$\sigma_Y = SD(Y) = \sqrt{np(1-p)}$

For large samples, the distribution of $\hat{P}$ is approximately Normal with mean $= p$ and standard deviation $= \sqrt{\frac{p(1-p)}{n}}$

Approximate Normality in large samples

Histogram of Bin (200, p=0.4) probabilities with superimposed Normal curve approximation. Recall that for Y~Bin(n,p). Y = # Heads in n-trials. Hence, the proportion of Heads is: Z=Y/n.

$\mu_Y = E(Y) = np$ $\qquad \mu_Z = E(Z) = \frac{1}{n}E(Y) = p$

$\sigma_Y = SD(Y) = \sqrt{np(1-p)}$ $\qquad \sigma_Z = SD(Z) = \frac{1}{n}SD(Y) = \sqrt{\frac{p(1-p)}{n}}$

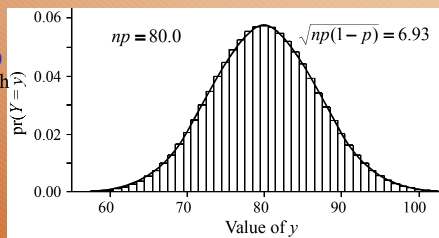
This gives us bounds on the variability of the sample proportion:

$$\mu_Z \pm 2SE(Z) = p \pm 2\sqrt{\frac{p(1-p)}{n}}$$

What is the variability of this proportion measure over multiple surveys?

Approximate Normality in large samples

Histogram of Bin (200, p=0.4) probabilities with superimposed Normal curve approximation. Recall that for Y~Bin(n,p)



The sample proportion Y/n can be approximated by normal distribution, by CLT, and this explains the tight fit between the observed histogram and a N(pn, $\sqrt{np(1-p)}$)

Standard error of the sample proportion

Standard error of the sample proportion:

$$se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Review

- We use both $\hat{p}$ and $\hat{P}$ to describe a sample proportion. For what purposes do we use the former and for what purposes do we use the latter? (observed values vs. RV)
- What two models were discussed in connection with investigating the distribution of $\hat{P}$? What assumptions are made by each model? (Number of units having a property from a large population Y~ Bin(n,p), when sample <10% of popul.; Y/n~Normal(m,s), since it's the avg. of all Head(1) and Tail(0) observations, when n-large).
- What is the standard deviation of a sample proportion obtained from a binomial experiment?

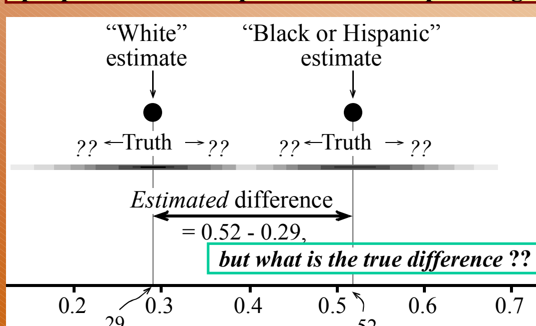
$$SD(Y/n) = \sqrt{\frac{p(1-p)}{n}}$$

Review

- Why is the standard deviation of $\hat{P}$ not useful in practice as a measure of the precision of the estimate? $SD(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$, in terms of p – unknown!
- How did we obtain a useful measure of precision, and what is it called? (SE($\hat{p}$))
- What can we say about the true value of p and the interval $\hat{p} \pm 2$ SE($\hat{p}$)? (Safe bet!)
- Under what conditions is the formula

SE($\hat{p}$) $= \sqrt{\hat{p}(1-\hat{p})/n}$ applicable? (Large samples)

Estimating a difference – proportions of people who believe police use racial profiling



Is there racial profiling or are there confounding explanatory effects?!?

- The book by Best (*Damned Lies and Statistics: Untangling Numbers from the Media, Politicians and Activists*, Joel Best) shows how we can test for racial bias in police arrests. Suppose we find that among 100 white and 100 black youths, 10 and 17, respectively, have experienced arrest. This may **look plainly discriminatory**. But suppose we then find that of the 80 middle-class white youths 4 have been arrested, and of the 50 middle-class black youths 2 arrested, whereas the corresponding numbers of lower-class white and black youths arrested are, respectively, 6 of 20 and 15 of 50. These arrest rates correspond to 5 per 100 for white and 4 per 100 for black middle-class youths, and 30 per 100 for both white and black lower-class youths. Now, better analyzed, the data suggest **effects of social class, not race as such**.

Standard error of a difference

Standard error for a difference between **independent** estimates:

$$SE(Est_1 - Est_2) = \sqrt{SE(Est_1)^2 + SE(Est_2)^2}$$

or

$$SE(\hat{\theta}_1 - \hat{\theta}_2) = \sqrt{SE(\hat{\theta}_1)^2 + SE(\hat{\theta}_2)^2}$$

Student's *t*-distribution

- For random samples from a Normal distribution,

$$T = \frac{(\bar{X} - \mu)}{SE(\bar{X})}$$

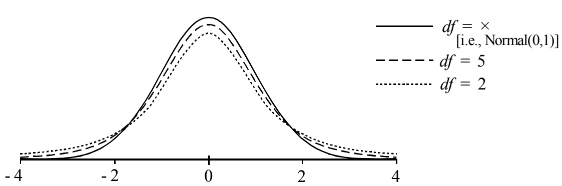
Recall that for samples from N(μ , σ)

$$Z = \frac{(\bar{X} - \mu)}{SD(\bar{X})} = \frac{(\bar{X} - \mu)}{\sigma / \sqrt{n}} \sim N(0,1)$$

is exactly distributed as Student($df = n - 1$) ← Approx/Exact Distributions

- but methods we shall base upon this distribution for T work well even for small samples sampled from distributions which are quite non-Normal.
- df is number of observations –1, degrees of freedom.

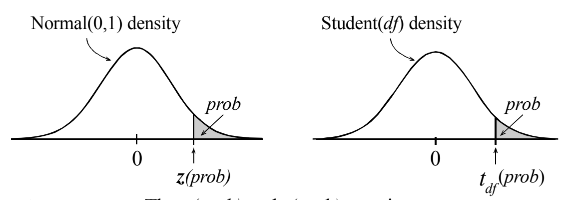
Density curves for Student's *t*



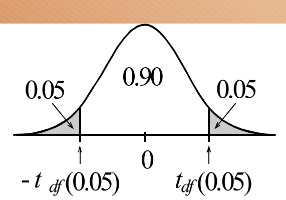
Student(df) density curves for various *df*.

Notation

- By $t_{df}(prob)$, we mean the number t such that when $T \sim$ Student(df), P($T \geq t_{df}$) = $prob$; that is, the tail area above t (that is to the right of t on the graph) is $prob$.

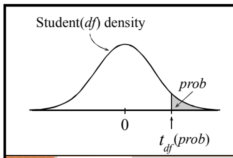


The *z(prob)* and *t(prob)* notations.



The central 90% of the Student(df) distribution.

Reading Student's *t* table



Desired *upper-tail prob*

Extracts from the Student's t-Distribution Table

prob

df	.20	.15	.10	.05	.025	.01	.005	.001	.0005	.0001
6	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959	8.025
7	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7.063
8	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041	6.442
...	...	...	...	...	...	...	...	...	...	...
10	0.879	1.093	1.372	1.812	$\boxed{2.228}$	2.764	3.169	4.144	4.587	5.694
...	...	...	...	...	...	...	...	...	...	...
15	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073	4.880
...										
∞	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.719

Desired *df*

t-value

Do we need an simulation of *T* and *Z* scores? Use the Online compute-engine …

Review

- Qualitatively, how does the Student (df) distribution differ from the standard Normal(0,1) distribution? What effect does increasing the value of df have on the shape of the distribution? (σ is replaced by SE)
- What is the relationship between the Student ($df=\infty$) distribution and the Normal(0,1) distribution? (Approximates N(0,1) as n→increases)

Review

- Why is T, the number of standard errors separating $\overline{X}$ and μ, a more variable quantity than Z, the number of standard deviations separating $\overline{X}$ and μ? (Since an additional source of variability is introduced in T, SE, not available in Z. E.g., P(-2<=T<=2)=0.9144 < 0.954=P(-2<=Z<2), hence tails of T are wider. To get 95% confidence for T we need to go out to +/-2.365).
- For large samples the true value of μ lies inside the interval $\overline{x} \pm 2$ se($\overline{x}$) for a little more than 95% of all samples taken. For small samples from a normal distribution, is the proportion of samples for which the true value of μ lies within the 2-standard-error interval smaller or bigger than 95%? Why?(Smaller – wider tail.)

Review

- For a small Normal sample, if you want an interval to contain the true value of μ for 95 % of samples taken, should you take more or fewer than two-standard errors on either side of $\overline{x}$? (more)
- Under what circumstances does mathematical theory show that the distribution of $T=(\overline{X}-\mu)/\mathrm{SE}(\overline{X})$ is exactly Student ($df=n-1$)? (Normal samples)
- Why would methods derived from the theory be of little practical use if they stopped working whenever the data was not normally distributed? (In practice, we're never sure of Normality of our sampling distribution).

Summary

Sampling distribution of $\overline{X}$

Sample mean, $\overline{X}$:

For a random sample of size n from a distribution for which $E(X)=\mu$ and $sd(X)=\sigma$, the sample mean $\overline{X}$ has :

- $\mathrm{E}(\overline{X}) = \mathrm{E}(X) = \mu, \quad \mathrm{SD}(\overline{X}) = \dfrac{\mathrm{SD}(X)}{\sqrt{n}} = \dfrac{\sigma}{\sqrt{n}}$
- If we are sampling from a Normal distribution, then $\overline{X} \sim$ Normal. (**exactly**)
- *Central Limit Theorem*: For almost any distribution, $\overline{X}$ is **approximately** Normally distributed in large samples.

Sampling distribution of the sample proportion

- **Sample proportion, $\hat{P}$:** For a random sample of size n from a population in which a proportion p have a characteristic of interest, we have the following results about the sample proportion with that characteristic:
  - $\mu_{\hat{p}} = \mathrm{E}(\hat{P}) = p; \qquad \sigma_{\hat{p}} = \mathrm{SD}(\hat{P}) = \sqrt{\dfrac{p(1-p)}{n}}$
  - $\hat{P}$ is approximately Normally distributed for large n

(e.g., $np(1-p) \geq 10$, though a more accurate rule is given in the next chapter)

Standard error

- **The standard error**, SE($\hat{\theta}$), for an estimate $\hat{\theta}$ is:
  - an estimate of the std dev. of the sampling distribution
  - a measure of the precision of $\hat{\theta}$ as an estimate of θ
- **For a mean**
  - The sample mean $\bar{x}$ is an unbiased estimate of the population mean μ
  - $\mathrm{SE}(\bar{x}) = \frac{s_X}{\sqrt{n}}$

Standard errors cont.

- **Proportions**
  - The sample proportion $\hat{p}$ is an unbiased estimate of the population proportion p
  - $se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- **Standard error of a difference:** For independent estimates,

$$\mathrm{se}(\hat{\theta}_1 - \hat{\theta}_2) = \sqrt{\mathrm{se}(\hat{\theta}_1)^2 + \mathrm{se}(\hat{\theta}_2)^2}$$

Some Parameters and Their Estimates

	Population(s) or Distributions(s) ↓ Parameters	Sample data ↓ Estimates	→ Measure of precision
Mean	m	$\bar{x}$	se $(\bar{x})$
Proportion	p	$\hat{p}$	se $(\hat{p})$
Difference in means	μ_1-μ_2	$\bar{x}_1 - \bar{x}_2$	se $(\bar{x}_1 - \bar{x}_2)$
Difference in proportions	p_1-p_2	$\hat{p}_1 - \hat{p}_2$	se $(\hat{p}_1 - \hat{p}_2)$
General case	θ	$\hat{\theta}$	se $(\hat{\theta})$

Student's *t*-distribution ….

- Is bell shaped and centered at zero like the Normal(0,1), but
- More variable (larger spread and fatter tails).
- As *df* becomes larger, the Student(*df*) distribution becomes more and more like the Normal(0,1) distribution.
- Student($df = \infty$) and Normal(0,1) are two ways of describing the same distribution.

Student's *t*-distribution cont.

- For random samples from a Normal distribution,

$$T = (\bar{X} - \mu) / SE(\bar{X})$$

  is exactly distributed as Student($df = n$ - 1), but methods we shall base upon this distribution for T work well even for small samples sampled from distributions which are quite non-Normal.
- By $t_{df}(prob)$, we mean the number t such that when $T \sim$ Student(df), pr($T \geq t$) = $prob$; that is, the tail area above t (that is to the right of t on the graph) is $prob$.