BMI 713 / GEN 212

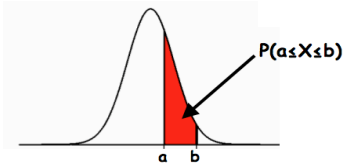
Lecture 2: Hypothesis Testing

- Normal and *t* distributions
- Inference on mean
- Hypothesis testing
- One-sample t-test

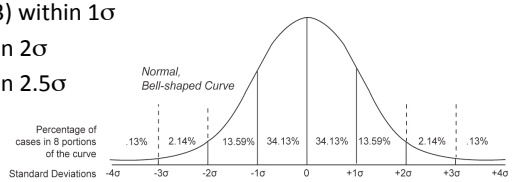
September 16, 2010

Normal Distribution

- $P(a \le X \le b)$ = probability that X falls between *a* and *b*



- 68% (~2/3) within 1σ
- 95% within 2σ
- 99% within 2.5σ



Standard Normal

- To figure out the probability that a normal random variable falls in a given range, first transform the variable into a standard normal variable.

"z-score"

- If $X \sim N(\mu,\sigma^2)$ and $Z = \dfrac{X-\mu}{\sigma}$, then $Z \sim N(0,1)$

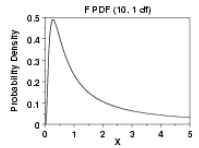
- In fact, this is the form of most statistical testing:

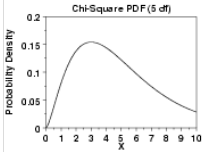
$$\frac{\text{Statistic - Hypothesized value}}{\text{Square root of the variance of the statistic}}$$

follows a known probability distribution

A Little Bit of Statistical Theory

- There are few well-known distributions in statistics
  - Normal distribution
  - t-distribution
  - F-distribution
  - χ^2(chi-square)-distribution
  - Binomial distribution (discrete)
  - Poisson distribution (discrete)





Example

- Systolic blood pressure in a population is normally distributed with mean 140 and std. dev. 9. What fraction of the population has SBP ≥155?"
- Solution: $X \sim N(140,9^2)$

$$P(X \geq 155) = P\left(\frac{X-140}{9} \geq \frac{155-140}{9}\right)$$

$$= P\left(Z \geq \frac{155-140}{9}\right)$$

$$P(Z \geq 1.67) = 0.0475$$

Inference on a Mean

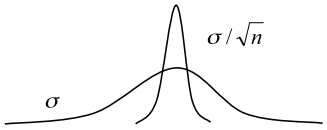
- Suppose height is normally distributed. I take a sample of 10 students and calculate the average. Is the deviation of this number from the population average unusual?
- Let's do a thought experiment
- Take a sample of 10, calculate its average $\bar{x}_1$
- Take another sample of 10, calculate its average $\bar{x}_2$
- Take another sample of 10, calculate its average $\bar{x}_3$
- What does the distribution of $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$ look like?
- What is distribution of the random variable $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$?

Distribution of Sample Mean

- $\bar{X}$ are approximately normally distributed when *n* is large

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

- The larger the sample size, the less spread you should see
- Standard deviation of sample mean → 'standard error'
- The denominator has square root of *n*, not *n*



Example

- Suppose the distribution of birth weights for infants whose gestational age is 40 weeks approximately normal with μ=3500 grams and σ=430 grams.
  - Given a newborn whose gestational age is 40 weeks, what is the probability that his or her birth weight is less than 2500 grams?

$$P(X < 2500) = P\left(\frac{X-3500}{430} < \frac{2500-3500}{430}\right)$$

$$= P(Z < -2.34) = 0.01$$

  - What value cuts off the lower 5% of the distribution?

$$X = (-1.645)(430) + 3500 = 2793$$

Example Cont'd

- What is the distribution of means of samples of size 5?

  It is approximately normally distributed with μ=3500, std dev $= \sigma/\sqrt{n} = 430/\sqrt{5} = 192$

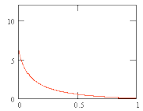
- Given a sample of five all with gestational age 40 weeks, what is the probability that their mean birth weight is less than 2500 grams?

$$P(\bar{X} < 2500) = P\left(\frac{\bar{X} - 3500}{192} < \frac{2500 - 3500}{192}\right)$$

$$= P(Z < -5.21) < 0.001$$

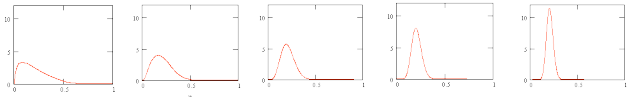
A Little Bit of Statistical Theory

- For $\bar{X}$ to be normally distributed, does X have to be normally distributed?
- The distribution of an average tends to be Normal, even when the distribution from which the average is computed is non-Normal (Central Limit Theorem)



Suppose my variable of interest has this distribution. What is the distribution of the averages?

n=2 n=4 n=8 n=16 n=32



Confidence Intervals

- Previously, we saw that for a normal variable, 95% of the data are contained in $(\mu - 2\sigma, \mu + 2\sigma)$
- So, we know that there is 95% probability that

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

  contains the true mean
- After we draw the sample we cannot say, "The probability that μ is contained in the interval is 95%."
- μ is fixed, not random. Once we have calculated the interval, it simply either contains μ or it doesn't.

Hypothesis Testing

- *Statistical inference:* drawing conclusions about an entire population based on the information in a sample
- Specify the null hypothesis (H_0) and the alternative hypothesis (H_1)
- Select a significance level and calculate the statistic
- Calculate the p-value (the probability of obtaining a statistic as extreme or more extreme under the null hypothesis)
- Describe the result and the conclusion in an understandable way
- You "fail to reject H_0" rather than "accept H_1"

One-sample Inference

- Null hypothesis: a statement that the population parameter is equal to some particular value of interest.

$$H_0 : \mu = \mu_0$$

- "Proof by contradiction": Null hypothesis is typically what we want to disprove.

Alternative Hypothesis

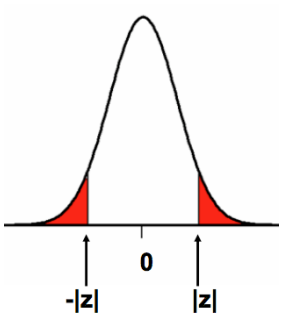
- For a given H_0, various alternatives are possible:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0 \qquad H_1 : \mu \neq \mu_0 \qquad H_1 : \mu > \mu_0$$

- Strategy: check whether the difference between the sample mean and the "null value" μ_0 is too big to be due to chance alone

P-value

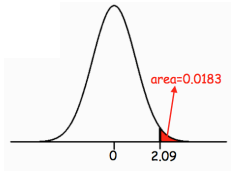
- p-value is the probability that a N(0,1) random variable would be greater than |z| or less than -|z|
- In hypothesis testing, p-value is the probability you'd get a sample estimate as extreme as the one you got (relative to μ_0) or more extreme, if H_0 were true.



Example

- Fasting plasma glucose levels are measured on a sample of 20 mice. The sample average is 107 mg/dL. Suppose that the standard deviation in this population is known to be 15. Is there evidence that this population has average FPG > 100 (i.e., impaired glucose tolerance)?

$$H_0 : \mu = 100$$
$$H_1 : \mu > 100$$
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{107 - 100}{15/\sqrt{20}} = 2.09$$
$$P(Z \geq 2.09) = 0.0183$$



P-value

- Small p-value → data would have been unlikely if H_0 were true, so reject H_0
- If H_0 is rejected, the result is "statistically significant"
- If H_0 is not rejected, it does not mean that H_0 is true.

- **"Not guilty" is not the same thing as "innocent"!**
- It is incorrect to talk about the "probability that H_0 is true" (or false). Either it's true or it's not --we just don't know. **Inference** means deciding whether to believe it's true or not.

P-value

- The smaller the p-value, the more convinced we are that it's real.
- **A small p-value does not mean that the difference between μ0 and the true value μ is large.**
- In other words, statistical significance measures whether a result is "real", not whether it's large
- Example:
  - With genomic data, it is easy to get a small p-value due to the large number of data points!
  - The correlation coefficient between two variables may only be .01, but it could still be statistically significant (p<.0001)

What if Variance is Unknown?

- What is the problem with the previous examples?
- We do not know the population σ.
- We estimate σ with the sample standard deviation s

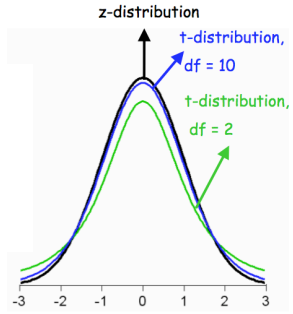
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- This introduces another source of uncertainty, so we must modify our hypothesis test to reflect that. This modification changes our "z-test"to a "t-test"

One sample t-test

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

t-test $$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$



- t-distribution has fatter tails than z-dist; more diffuse distribution reflects greater uncertainty
- For large *n*, t_{n-1} is nearly identical to Normal

A Little Bit of Statistical Theory

- These are *asymptotic* results
- That means the result (e.g., p-value) becomes more accurate as the sample size gets large
- How big should my sample size be? How quickly does it become valid?
- It depends on the underlying distribution: if the underlying distribution is normal, then you do not need as many samples
- Most text books will give you guidelines

- What if my sample size is still small?

When the sample size is small

- What is the problem with the t-test in this case?
  - You may have an inaccurate estimate of the variance

- Example from genome-wide gene expression analysis
  - The top genes might be those for which the variance was underestimated due to small sample size
- "Regularized t-test"
  - One solution using a fudge factor s_0 (Tusher et al, *PNAS*, 2003)

$$t = \frac{\overline{X} - \mu}{s_0 + s/\sqrt{n}}$$

Summary

- Normal distribution
- Sample distribution of a mean
- Null hypothesis, p-value
- z-test; t-test

- There are many assumptions behind the tests
  - Distributional assumption
  - minimum sample size
- Recognize that every statistic has flaws--the question is whether it is severe enough to invalidate the conclusion
- Consult a statistician for help but recognize that not all statisticians are the same